7/

# THE JOURNAL OF PHILOSOPHY

CUK-H05966-13-KP1374

UNIVERSITY

16197

13

VOLUME XCIX, NUMBER 7
JULY 2002

*page*
333 *What Vague Objects Are Like* Michael Morreau
362 *Mathematics and Metaphilosophy* Jerrold J. Katz

390 NOTES AND NEWS

*Published by* **The Journal of Philosophy, Inc.**

# THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

*Editors*: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

You may also visit our website at: www.journalofphilosophy.org

K.P. 1374
dt. 05/8/08

# THE JOURNAL OF PHILOSOPHY

2002

SUBSCRIPTIONS (12 issues)

| | |
|---|---|
| Individuals | $35.00 |
| Libraries and Institutions | $75.00 |
| Students, retired/unemployed philosophers | $20.00 |
| Postage outside the U.S. | $15.00 |

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Ephrata, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by writing to us

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices
POSTMASTER: Send address changes to the *Journal of Philosophy* at MC 4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2002 by the Journal of Philosophy, Inc. ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY
VOLUME XCIX, NO. 7, JULY 2002

16/9/02
-1- 8f Philo

# WHAT VAGUE OBJECTS ARE LIKE*

We cannot precisely trace the outline of a cat, the limits of a city, or the edges of a cloud—not right down to the last tiny detail. Might ordinary things such as these lack sharp boundaries and in this sense really be vague?

Many think it incomprehensible that ordinary things might be vague. Vague objects, they think, must have some sort of shady presence, being neither quite there nor not there, and they must have dodgy identities, being somehow neither quite the same as other things nor different from them. Vague objects, if so, cannot be at all ordinary. On the contrary, they really must be pretty queer.

In fact, we can take the hypothesis that ordinary things are vague quite seriously. It is coherent, since vague objects need not have a shady presence or dodgy identities. It is plausible, since sometimes there seems to be no matter of fact whether one object is a part of another. It is even fairly conservative, since, surprisingly, indefinite parthood is compatible with received ideas about parts and wholes, as expressed by classical mereology. We ought not to pass up this hypothesis too quickly. It might be our best shot at preserving sensible, everyday ideas about what there is in the world.

I proceed as follows. Section I contrasts the hypothesis that there are vague objects with the orthodoxy that vagueness is simply a matter of thought and language. Section II explains why many have thought that vague objects must have a shady presence and dodgy identities, and why it is wrong to think so. Section III argues that ordinary objects are vague because there is sometimes no saying whether one

---

* With thanks for helpful comments to Dalia Drai, Terence Horgan, Dick de Jongh, Jerry Levinson, Jonathan Lowe, Paul Pietroski, Carl Posy, Theodore Sider, and Timothy Williamson.

0022-362X/02/9907/333–61

of them is a part of another. Section IV motivates the introduction, in section V, of a class of abstractions, fuzzy boundaries. The final section VI uses fuzzy boundaries to construct a world of vague abstract objects. These objects are fully present and they have determinate identities. Their mereology is classical.

## I VAGUE OBJECTS AND REPRESENTATIONS

Take Tibbles, the cat. The one loose whisker used to be firmly attached and soon it will drop off for good; meanwhile, it is a *questionable* part of Tibbles. It is neither definitely a part nor definitely not a part of her. Or take the city of Toronto. The tree in the outskirts is a questionable part and so the city, like the cat, has no sharp boundary. Or take just about any ordinary material object. Look closely and there is always something tearing away or coming loose. Look even more closely and there are always microscopic particles wearing off at the edges or evaporating away. The world seems to be full of objects that lack sharp boundaries and in this sense really are vague.[1]

Even so, there might be no vague objects. Instead of a vague cat there might just be many precise quantities of animal tissue that almost completely overlap, only differing around the edges by the odd whisker or hair. In this case, each of these precise quantities has a claim to be called 'Tibbles', but none has a special claim, and this name fails to pick out any one candidate in particular. Questionable parts of Tibbles are then just things that are parts of some candidates but not others. Similarly, there might just be many precisely delimited localities, equally deserving to be called 'Toronto', some including the questionable tree, some not. In this case, the *names* 'Toronto' and 'Tibbles' are vague, in the sense that they fail to pick out any one thing. But there is no vague city, and no vague cat.

Orthodoxy has it that there are indeed no vague objects. Words, thoughts, pictures, and other representations are vague, but apart from them the world is precise. Bertrand Russell[2] was the first to say so. He argued that to think the world is vague is to mistake properties of thoughts for properties of things, and blamed idealism for cultivating confused habits of thinking that can lead to this sort of mix-up. One might as well think, he scoffed, that the world is also muddleheaded. There is no mistaking the appeal of this idea. Certainly, the relation between representations and what they represent is somewhat indefinite; and if this fact by itself can explain the phenomena of vagueness, then it might seem idle to imagine any other vagueness in the world. It might seem simpler and better to suppose that, apart from their representations in thought and language, all things are precise.

---

[1] I use 'object' broadly to cover organisms, cities, and abstract entities, as well as ordinary material objects.

[2] "Vagueness," *Australian Journal of Philosophy and Psychology*, I (1923): 84-92; reprinted in Rosanna Keefe and Peter Smith, eds., *Vagueness: A Reader* (Cambridge: MIT, 1997), pp. 61-68.

But orthodoxy is hard put to explain some of the phenomena. Take an ordinary material object, say Tibbles. If vagueness is all a matter of representation, there is no vague cat. There are just the many precise cat candidates that differ around the edges by the odd whisker or hair. Since there is a cat, Tibbles, and since orthodoxy leaves nothing else for her to be, one of these cat candidates must then be a cat. But if any is a cat, then also the next one must be a cat; so small are the differences between them. So *all* the cat candidates must be cats. The levelheaded idea that vagueness is a matter of representation seems to entail that wherever there is a cat, there are a thousand and one of them, all prowling about in lockstep or curled up together on the mat. That is absurd. Cats and other ordinary things sometimes come and go one at a time.[3]

Someone might try to contain the outbreak of cats by explaining that none of the candidates is a cat, but that is not a levelheaded thing to do. It follows that there is no cat, not one, and by implication that there are no ordinary things at all in the world: no babies, no bath water, nothing. Common sense balks.[4] One might instead explain that, although there are many different candidates, they are all one and the same cat.[5] Or that just one of them is a cat but there is no saying which, or that although all the candidates are different cats, this is no matter since, so very similar, they are *almost* one and the same cat.[6] These explanations do not fly in the face of common sense; but there are many of them, and none is obviously right. You start to wonder, with all this explaining away that orthodoxy has to do, whether the hypothesis that ordinary objects are vague might square better with sensible ideas about what there is in the world.[7]

---

[3] There are versions of this puzzle in Peter Unger, "The Problem of the Many," *Midwest Studies in Philosophy*, v (1980): 411-68; and in Peter Geach, *Reference and Generality*, third edition (Ithaca: Cornell, 1980), p. 215.

[4] Unger does not balk: "There Are No Ordinary Things," *Synthese*, XLI (1979): 117-54.

[5] Geach illustrates the notion of relative identity with his version of the problem of the many cats (*op. cit.*).

[6] As David Lewis explains in "Many, But Almost One," in Keith Campbell, John Bacon, and Lloyd Reinhardt, eds., *Ontology, Causality, and Mind. Essays on the Philosophy of D.M. Armstrong* (New York: Cambridge, 1993), pp. 23-42.

[7] Others who think that various sorts of things might really be vague include Arthur W. Burks, "Empiricism and Vagueness," this JOURNAL, XLIII, 18 (August 29, 1946): 477-86; Kit Fine, "Vagueness, Truth and Logic," *Synthese*, XXX (1975): 265-

## II VAGUE EXISTENCE AND IDENTITY

The orthodoxy that vagueness is all a matter of thought and language has long been held in place by the belief that there is no real alternative. The idea that ordinary objects might really be vague has seemed too mysterious to take seriously. In part, this is because it has been thought that vague objects could not be fully present or have determinate identities. I shall show that there is no good reason to think so.

The misconception that vague objects cannot be fully present seems to be rooted in the notion that, if things are to have questionable parts, then *composition* must be vague, in the sense that there is sometimes no matter of fact whether some things make up a whole. Classically, a part of a whole is something that composes with some other things to make up this whole. So, one might think, if ever there is no matter of fact whether one thing is a part of another, that must be because there is no matter of fact whether the questionable part and some other things compose to make it up.[8] But then, one might think, there can be no matter of fact whether something, the composition of the questionable part and these other things, exists.[9] Now *this* idea is genuinely mysterious. How can something neither quite be nor not be there? Must we imagine that the presence of vague objects is somehow a matter of degree, like the intensity of a beam of light?

There is no need, for such thinking is mistaken. In fact, there can be vague objects though composition is precise. Imagine a vague cloud that has as its questionable part a wisp of water vapor on the edge. Suppose that, largely overlapping with this cloud, there is a collection of water droplets—cloud-minor—that is just like the cloud except that the wisp of vapor is a definite nonpart of it. Since the wisp is a questionable part of the cloud, we can suppose furthermore that

---

300, footnote 10; Mark Sainsbury, "Concepts without Boundaries," in Keefe and Smith, eds., pp. 251-64; Michael Tye, "Vague Objects," *Mind*, XCIX (1990): 535-57; John A. Burgess, "Vague Objects and Indefinite Identity," *Philosophical Studies*, LIX (1990): 263-87; Peter van Inwagen, *Material Beings* (Ithaca: Cornell, 1990); Eddy Zemach, "Vague Objects," *Noûs*, XXV (1991): 323-40; Richard Heck, "That There Might Be Vague Objects (So Far as Concerns Logic)," *The Monist*, LXXXI, 2 (1998): 274-96; and Michael Dummett, "Is Time a Continuum of Instants?" *Philosophy*, LXXV (2000): 497-515.

[8] Vann McGee seems to have thought so. He claims that an argument for unrestricted composition "presupposes" that there is no vagueness in the world; see his "Kilimanjaro," *The Canadian Journal of Philosophy*, XXIII (1998): 141-63, p. 163. van Inwagen argues that questionable parthood and vague composition go together (*op. cit.*, p. 228). Lewis assumes that unrestricted composition is at odds with worldly vagueness—*On the Plurality of Worlds* (New York: Blackwell, 1986). Having rejected the idea of vague composition, Lewis claims that the only intelligible account locates vagueness in thought and language (p. 212).

[9] Compare van Inwagen, pp. 272-73.

there is no matter of fact whether the wisp composes with cloud-minor to make up the cloud. The crucial point is that they can still compose to make up *something*. Their composition just has to be something other than the cloud: it has to be something of which the wisp is not a questionable part, but rather a definite part. The suggestion is that questionable parthood need not restrict composition at all, and the abstract vague world of section VI bears it out. Some things within this world have questionable parts, but there is always a matter of fact whether any given things compose since they always do. Composition is completely unrestricted, and nothing has any sort of shady presence.

A second misunderstanding about vague objects has been the idea that there can be no saying which of them are which, or how many there are. Their very *identities* must be vague. This idea led Russell to brush off the hypothesis that there are vague objects with the remark that "things are what they are, and there is an end of it" (*op. cit.*, p. 62), and it led Gareth Evans[10] to argue more fully that since the identities of things are always definite, the hypothesis that there are "objects about which it is a *fact* that they have fuzzy boundaries" is incoherent. The objects of the vague world in section VI will refute this argument. Not only their presence but also their identities are quite definite. But this matter has been so very prejudicial that it ought not to hang over the discussion until the end. In the rest of this section, I shall show that the argument from definite identities is inconclusive.

I shall not call into question what I take to be the insight at the center of both Russell's remark and Evans's argument, namely, that there is always a fact of the matter whether any given things are identical. That there might sometimes be no fact of *this* matter really *is* hard to imagine. Things are identical to themselves but different from other things, and it is hard to see how, somehow between itself and everything else, there can be anything left for a vague object to be indefinitely identical to. Talk about vague objects with indefinite identities really does not appear to be talk about anything at all.

Antirealists, thinking truth is bound up with methods for finding out, might disagree. For instance, a mathematical intuitionist might think there is sometimes no matter of fact about the identities of choice sequences, conceived as entities that are not fully determinate in advance.[11] Perhaps, despite appearances and arguments, the idea of indefinite identities is at least coherent.[12] Be this as it may, we shall have no use for it here. For a long time, the suspicion has lingered that the idea of vague objects has something to do with idealism. I shall argue that ordinary material objects can be vague and, although my argument does not require it, to ward off suspicion that some sort of antirealism is involved we shall do well to rule out indefinite identities from the start. This is, of course, not to say that identity *statements* are always either true or false. They will still have truth-value gaps, as they must, when singular terms do not determinately denote.

---

[10] "Can There Be Vague Objects?" *Analysis*, XXXVIII (1978): 208. See also Nathan Salmon, *Reference and Essence* (Princeton: University Press, 1981), pp. 243-46.

The main problem with the argument from definite identities is just that there is no reason to think that things with fuzzy boundaries must have indefinite identities. Strangely, Evans did not even try to show that they must; perhaps it did not occur to him that having a fuzzy boundary and having an indefinite identity might be different things.[13] However this may be, the omission hides a crucial difficulty with his argument, which comes to light as soon as we try to complete it.

To establish that fuzzy boundaries make for indefinite identities, someone might reason as follows. Comparing a vague Tibbles to a precise Tibbles-candidate, we see that they exactly agree as far as the definite parts and nonparts of Tibbles are concerned. So any difference between them must be among the questionable whiskers and hairs. But then, since there is no saying whether these are parts of Tibbles, there is no saying whether there is any difference at all between Tibbles and the candidate. In other words, there is no matter of fact whether Tibbles and the candidate are one and the same.

This reasoning overlooks something. If vagueness is a real characteristic of things, in particular of cats, then the possible differences between Tibbles and the Tibbles-candidate are not confined to the questionable whiskers and hairs. In fact, there is an obvious difference between them after all: Tibbles is vague but the candidate is not vague. There can be no question whether they are identical, for they definitely are two different things, not one, and nothing in between.[14] Evidently, we have completed Evans's argument only by setting aside the possibility that the vagueness of things is real. Since the reality of their vagueness is just what is at issue, this argument, so completed, begs the question.

---

[11] L.E.J. Brouwer described the construction by means of choice sequences of what one might take to be indefinitely equal real numbers—"Über Definitionsbereiche von Functionen," *Mathematische Annalen*, XCVII (1927): 60-75.

[12] As Terence Parsons and Peter Woodruff argue in "Worldly Indeterminacy of Identity," *Proceedings of the Aristotelian Society*, XCV (1995): 171-91

[13] Sainsbury distinguishes between them in "What Is a Vague Object?" *Analysis*, XLIX, 2 (1989): 99-112.

[14] Leibniz's law entails that if $o=p$, then whatever is true of $o$ is also true of $p$. Here I use the contrapositive, which might seem unjustified. Truth must pass from the

We need not take Evans's argument to be incomplete. Instead we can take it as an argument that nothing is vague in the special sense of being indefinitely identical to something. Some have asked whether objects can be vague in this sense,[15] and although Evans speaks of objects with "fuzzy boundaries," perhaps this is what he too had in mind. Be this as it may, his argument, so construed, remains inconclusive. For it does not even touch the question whether things can be vague in the core sense that they lack sharp boundaries.

In summary, Russell was right that things are what they are. Things are identical to themselves, different from other things, and indefinitely identical to nothing. Still, the world might be vague since, for all anyone has said, also things without sharp boundaries can have definite identities. Also they can be "what they are." We have no good reason to think that, since identities are definite, the world must be precise.

## III. QUESTIONABLE PARTS

I shall now argue that some ordinary material objects are vague because they have questionable parts. The main task will be to say how parthood can sometimes be questionable. First, though, I shall contrast ordinary material objects with certain special objects that apparently cannot have questionable parts and are not vague. I have in mind objects picked out by expressions such as 'that Styrofoam' and 'the soap in China'. Call them *quantities* of matter.[16]

We can get an intuitive hold on questionable parthood by thinking of parts that have come loose and will be lost, such as the cat's whisker. But this will not help us to grasp the notion of questionable inclusion among quantities of matter, because these cannot gain or lose parts. Imagine taking a quantity of soap and then separating some of it from the rest. Afterward, the soap that you began with will still include the separate part; it will still include it even if you send the separate part somewhere

---

antecedent of an instance of Leibniz's law to the consequent, but where there can be truth-value gaps perhaps falsity need not pass the other way: arguably, the consequent might be false while the antecedent has a truth-value gap. Crucially, no one can say so in defense of Evans's argument. It uses the contrapositive of Leibniz's law to establish that identities are always definite (*op. cit.*).

[15] For example, Francis Jeffrey Pelletier, in "Another Argument against Vague Objects," this JOURNAL, LXXXVI, 9 (September 1989). 481-92.

[16] See Helen Morris Cartwright, "Quantities," *The Philosophical Review*, LXXIX (1970): 25-42. See also Tyler Burge, "Truth and Mass Terms," this JOURNAL, LXIX, 10 (May 18, 1972): 263-82, and Dean Zimmerman, "Theories of Masses and Problems of Constitution," *The Philosophical Review*, CIV (1995): 53-110.

else altogether, in which case some of the soap will end up there and some will remain here. You can *scatter* a quantity of something but, short of destroying it, you cannot keep some of it from being included in all of it. Quantities of matter have their parts essentially.

This is one reason to think there must be more to part-whole relations among material objects than just inclusion among quantities of matter. Unless ordinary thinking is very mistaken, when a whisker works loose and drops off a cat, the cat continues to exist but the whisker ceases to be a part of it. Organisms and other ordinary material objects have temporary parts. They do not have their parts essentially.

Another reason to think there must be more to part-whole relations than material inclusion is that not all included quantities count as parts. Among the parts of a cat we count a whisker, say, and the cat's nervous system, but we do not count the material content of an arbitrarily marked-off region within the cat. English makes a handy distinction. An arbitrary portion of cat is "some of" or "part of" the cat, its stuff included in the stuff of the cat. But it is not "*a* part of" the cat, in same full sense in which a firmly attached whisker is, or the nervous system.[17] Whether a thing is a part of the cat, in the relevant full sense, is not a simple matter of whether its stuff is some of cat's stuff. It depends on this thing's empirical properties.

I shall now argue that organisms stand in functional part-whole relations to some of their parts. Such relations are temporary and empirical and, I shall argue, sometimes there is no matter of fact about whether something is a functional part of an organism. This is one way in which it can happen that organisms lack sharp boundaries.

A *functional part* of the cat is something that is a part of the cat because it has a function that contributes to the functioning of the cat, and is furthermore suitably integrated into the rest of the cat for fulfilling this function. Functions, here, are *proper* functions of the sort that interest functional anatomists and reverse engineers; intuitively speaking, they are what things really are there for. For instance, it is a proper function of the cat's whisker to detect objects close to the cat's head. While the whisker is firmly attached it is suitably integrated for fulfilling this function and is a functional part of the cat. An arbitrary portion of the cat, on the other hand, has no proper function. It does have various capacities; for instance, it can make a meal for a dog. But an arbitrary portion of cat is not really *there* for anything like that. It is not a functional part of the cat.

---

[17] Richard Sharvy notes these differences between the relation expressed by 'some of' and by 'part of', and on the other hand that expressed by 'a part of'—"Mixtures," *Philosophy and Phenomenological Research*, XLIV (1983): 227-39; see section 3.

It might appear that to ascribe proper functions to naturally occurring things is to mistake properties of artifacts for properties of things that were not made for any purpose. What natural distinction is there to be drawn between proper functions and capacities in general? At issue is, in the end, whether there really are organisms, naturally occurring beings with a proper-functional organization. I shall not settle this question here. Notice, though, that evolutionary history seems to draw a line separating proper functions of biological parts from capacities in general. The capacity of whiskers to detect objects once contributed to the fitness of the ancestors of cats. It helped them to survive and reproduce. The capacity of arbitrary portions of their bodies to feed ancestral dogs, on the other hand, did not contribute to their fitness. Plausibly, there is in their etiology a natural basis for ascribing proper functions to biological parts.[18]

Crucially, for my present purpose, organisms are always gaining and losing functional parts. Once the loose whisker was firmly attached to the rest of the cat. Then it was suitably integrated for detecting objects and it was a functional part of the cat. Soon, when it has dropped off, it will no longer be suitably integrated and it will no longer be a part of the cat. This is why organisms are vague. They gain and lose their functional parts *gradually*. There is no precise moment at which the whisker becomes so loose that it can no longer fulfill its function. It does not go in an instant from being a part of the cat to being a nonpart but is, for a time, a questionable part of the cat. Indeed organisms, maintaining themselves in continuous flux, are more vague than meets the eye. As a flake of the cat's skin loosens, it does not instantly become so loose that it can no longer do its bit to keep the inside environment of the cat separate from its external environment. It too takes a while to quit its small part in the life of the cat and is, for a while, a questionable part of the cat.

Others, too, have given reasons to think that biological systems have questionable parts. Robert Cummins,[19] with a somewhat broader conception of function in mind, has argued that what makes something a part of the nervous system is that it has capacities that figure in a functional analysis of the capacity to respond to external stimuli, to coordinate movement, and so on. As Cummins points out, if this is what makes something a part of the nervous system, then it is questionable whether, say, the glial cells are a part of it (*ibid.*, footnote 18).

---

[18] Those who have tried to ground proper function in etiology include Ruth Millikan, in "Biosemantics," this JOURNAL, LXXXVI, 6 (June 1989): 281-97; and Karen Neander, in "Functions as Selected Effects: The Conceptual Analyst's Defense," *Philosophy of Science*, LVIII (1991): 168-84. Paul Griffiths extends their etiological conception of proper function from organisms to artifacts in "Functional Analysis and Proper Functions," *British Journal for the Philosophy of Science*, XLIV (1993): 409-22.

[19] "Functional Analysis," this JOURNAL, LXXII, 20 (November 20, 1975): 741-65.

How can material objects be vague if, as might appear to be the case, quantities of matter cannot have questionable parts and are precise? Not, we have assumed, by being indefinitely identical to quantities of matter. There is another way, though. Suppose there is a special relation of *constitution*, distinct from identity, in which, say, a statue stands to a quantity of bronze, or a cat to a quantity of animal tissue.[20] Then material objects can be vague if they are indefinitely constituted by quantities of matter without being indefinitely identical to them. Notice the formal parallel between the vagueness of representations and the vagueness of indefinitely constituted objects: the name 'Tibbles' is vague if there is no matter of fact which of several candidates it picks out; the cat Tibbles is vague if, having questionable parts, there is no matter of fact which of several candidates constitutes her.[21] If neither quantities of matter nor identities are vague, the idea that ordinary material objects are vague is about as good as the idea that they are constituted by quantities of matter without being identical to them.

Just *how* good this is depends on something else. The distinction between constitution and identity goes with the doctrine that ordinary material things are continuants, persisting through time and meanwhile undergoing changes. It is with this distinction that one can explain how it is that a continuant cat is made up of different quantities of matter at different times. Someone who thinks ordinary material objects are continuants ought to have no special difficulty with the idea that some of them are vague. The alternative is that material objects persist through time only in the improper sense that, like processes and events, they have earlier and later temporal parts.[22] Someone who thinks that material objects have temporal parts will not have the same reason to distinguish constitution from identity, for he can explain that an object has a variable makeup insofar as it is *identical-for-a-while* to each of several different quantities of matter, having a different temporal part in common with each one. He is unlikely to think that material objects can be vague. To him it will seem that they must be just as precise as the different quantities of matter to which they are, for a while, identical.

---

[20] Those who separate constitution from identity include Burge (*op. cit.*, p. 278) and "Mass Terms, Count Nouns and Change," *Synthese*, xxxi (1975): 459-78; David Wiggins, *Sameness and Substance* (Cambridge: Harvard, 1980), pp. 30-35; Jonathan Lowe, "The Problem of the Many and the Vagueness of Constitution," *Analysis*, xlviii (1982): 27-30; Mark Johnston, "Constitution Is Not Identity," *Mind*, ci (1992): 89-105; and Zimmerman.

[21] Vague constitution has been proposed before as a solution to the problem of the many. See Lowe; and Johnston, p. 101.

[22] See, for instance, Nelson Goodman and W. V. Quine, "Steps Toward a Constructive Nominalism," *The Journal of Symbolic Logic*, xii (1947): 105-22; see also Quine, *Word and Object* (Cambridge: MIT, 1960), p. 171.

In conclusion, the hypothesis that ordinary material objects are vague fits naturally within the doctrine that they are three-dimensional continuants. Someone who thinks ordinary material objects have temporal parts can still accept this hypothesis if he distinguishes between constitution and identity, but he will lack the three-dimensionalist's independent reason to distinguish between them. He can also accept this hypothesis if he thinks that there are vague quantities of matter or that identity can be indefinite. Then he can say that some objects are identical-for-a-while to vague quantities, or vaguely-identical-for-a-while to precise ones. While the idea of vague quantities of matter appears to be coherent, I know of no independent reason to think there are any such things. The notion of vague identities, on the other hand, defies comprehension.

Let us now put organisms to one side, turning instead to social entities such as cities, nations, clubs, and families. I shall argue that they can be vague because of indeterminacy in the relations of *belonging* in which a city stands to its environs and a nation to its territories, and in the relations of *membership* in which a club or family stands to the people that make it up. Interestingly, we speak of belonging, membership, and parthood in much the same terms. The Cook Islands "belong to" New Zealand or else they are "a part of" that nation; someone "belongs to" a family or club or else he is "a member of" or "a part of" it. Without going into how these matters are related outside of idiom, I shall also speak of them in the same terms.

Belonging is sometimes questionable because it is established on the balance of prima facie claims. If you happen to find some trivial item then the fact that you did makes it yours—unless someone else finds it at the same moment, in which case the two of you will have competing claims and it will be questionable whether the item is yours. Similarly, what made the Cook Islands a part of New Zealand was a legal act of annexation, undertaken after petitioning by the chiefs of the larger islands. As it happened, parthood was duly established on the balance of legal reasons, but questionable parthood could have resulted had there been a conflict of reasons. We can imagine a case. Suppose that, having quarreled, half of the chiefs had petitioned New Zealand for annexation but the other half Australia, and suppose that each country had proceeded in a way that, had it proceeded alone, would have resulted in its having sovereignty over the Cook Islands. Then there would have been conflicting claims to sovereignty. The Cook Islands would have been a questionable part of New Zealand and also a questionable part of Australia.

This explains how organisms, countries, and other things can lack sharp boundaries. In the following section, I shall distinguish among varieties of vague things according to how deeply their vagueness is compounded, and how deeply it is embedded within their part-whole structure. I shall also introduce the notion of a sharpening of a vague object. This will motivate the introduction of fuzzy boundaries in section v, and the construction of an abstract vague world in section vi.

## IV VARIETIES OF VAGUE OBJECTS AND THEIR SHARPENINGS

Natural change is continuous. Just as there is no precise moment at which the loose whisker becomes a nonpart of the cat, there is none at which it becomes a questionable part, either. As well as a first-order penumbra of questionable parts, in limbo between parthood and nonparthood, a vague object can have second-order penumbrae of objects that *questionably* are questionable parts, in limbo either between parthood and questionable parthood, or between questionable parthood and nonparthood. There can be third-order penumbrae of objects in limbo between penumbrae of lower orders, and so on. Ranking vague objects according to their highest-order penumbrae, we can distinguish vague objects of various orders of vagueness.

The continuous nature of change is one source of higher-order vagueness. Where parthood is established on the balance of reasons, as it seems to be in the case of countries, and where questionable parthood is the result of conflicts between reasons, conflicts between higher-order reasons are a further source of higher-order vagueness. Suppose there are secondary reasons that can resolve conflicts between primary reasons, and with them questions about parthood, by excluding primary reasons from the balance of reasons.[23] In this case, whether there is a conflict between primary reasons and whether one thing is a questionable part of another will turn on which primary reasons go on the balance. Where higher-order reasons conflict among themselves it can be questionable which ones go on the balance, and therefore questionable whether one thing is a questionable part of another. We might expect the orders of vagueness coming from conflicts between legal reasons, like the orders of the reasons themselves, to have a finite upper bound. Legal reasons are made, not found, and so far we can have made only so many of them.

---

[23] Joseph Raz introduced "exclusionary reasons" in *Practical Reason and Norms* (London: Hutchinson, 1975).

In the remainder of this section, I shall develop a distinction that can be drawn among vague objects with different levels of parts. It has to do with how deeply vagueness is buried within their part-whole structure.

Some objects seem naturally to break down into a top level of parts, when notionally taken apart. They seem to have *immediate* parts. Take, for instance, the cat's nervous system. Arguably it is an immediate functional part of the cat, a functional part of the cat that makes its contribution to the life of the cat directly, and not by being a functional part of another entity intermediate between it and the cat. The brain, on the other hand, a functional part of the nervous system, makes its contribution indirectly, and is not an immediate functional part of the cat.

Let the *penumbral* parts of a vague object be those things whose parthood status is somehow unsettled: they are questionable parts of it, or questionably questionable parts, or they are in some other sort of mereological limbo. Along with its immediate parts, we can assign to an object's top level of parts any *immediate penumbral* parts that it has. These are penumbral parts that, intuitively speaking, have what it takes to be immediate parts except that they are not quite parts. They are penumbral parts that *would* be immediate parts, if only they were parts.

Of course, there are more ways than one to carve up a cat. Perhaps there is no saying once and for all which are the immediate parts of things, and which are their immediate penumbral parts, but rather a range of different ways, depending on our concerns and interests. Anatomy texts break down higher organisms first into their main organ systems, but that need not keep us, in an everyday context, from thinking of Tibbles as immediately made up of her head, torso, and tail. In the following section, vague objects will have their fuzzy boundaries relative to a specification of their immediate (penumbral) parts, so this contextuality might appear to undermine the idea that objects can really be vague, independently of how we represent them to ourselves. It does not, though. For one thing, objects with a proper-functional organization really can have immediate parts, independently of our thinking. They really can have their fuzzy boundaries. Furthermore, my main point is not that anything *is* vague. That is, in part, an empirical matter. My main point is that the idea of worldly vagueness is coherent: things *can* be vague. Grant me, for the sake of the argument, that things can have immediate parts. This opens up one way—surely not the only one—toward a coherent and not unrealistic conception of vague objects.

Now we can distinguish among vague objects, according to where vagueness appears within their part-whole structures. An object is *immediately vague* if it has immediate penumbral parts. Such an object is vague right at the top level of parts. Otherwise, an object is *immediately precise*. Any vagueness within it is to be found only at lower levels of the part-whole structure.

Organisms tend to be immediately precise. Take Tibbles. Her head is firmly stuck on and so are her paws and tail. She does have penumbral parts, but they are just whiskers, hairs, molecules, and other things further down the part-whole structure. None of her penumbral parts would be, were it a part of Tibbles, a top-level part. She has no immediate penumbral parts. Tibbles, like any happy cat, is immediately precise. Immediate vagueness—with head, paws, or tail in some sort of mereological limbo—is a sorry state for a cat.

Countries also tend to be immediately precise. At the borders of a country, there might be questionable clods of earth and clumps of grass, but there is nothing that, were the borders sharpened up, could be counted among major geographical regions, infrastructure systems, and so on as an immediate part of the country. There are no immediate penumbral parts. In the imaginary case of a territorial dispute about the Cook Islands, on the other hand, New Zealand is immediately vague. Among the questionable parts of this nation there is a major geographical region, the Cook Islands, which would be an immediate part, if it were a part. Immediate vagueness is bad for countries as well as for cats. They fight wars to resolve it.

This brings us to the notion of a *sharpening* of a vague object, which we shall find useful in the next section. The penumbral parts of a vague object have a more or less unsettled status, in limbo between penumbrae of lower orders. Now take an immediately vague object and consider those of its immediate penumbral parts that have the least-settled status, the ones in the highest-order penumbrae. We can imagine settling the status of these objects somewhat, by distributing them among suitable penumbrae of lower orders. There will, in general, be many ways to do this, and for each way we can imagine an *immediate sharpening* of the vague object in question, an object just like it except that the status of the relevant penumbral parts *is* somewhat more settled, and in just this way. The immediate sharpening of a vague object is less vague than it is, since any vagueness left at the top level of parts is of a lower order of vagueness. There may be no vagueness at all left at the top level. Then the immediate sharpening is immediately precise.

So far, only immediately vague objects have sharpenings, but soon other vague objects will have them, too. One might object that each sharpening of a vague cat will have an equal claim to be called a 'cat', and that along with sharpenings comes a new problem of the many. Not so—unlike the earlier cat candidates, the many sharpenings of a vague object do not overlap. They are not all together in the same place. The vague object *could* have had clearer relations to its penumbral parts; and had the world been slightly different—had the whisker been more firmly attached, had the chiefs been less quarrelsome—it *would* have had clearer relations to them. A sharpening of a vague object is a possible object that *does* have clearer relations to them. The vague cat is here in this possible world. Her many sharpenings are elsewhere, in as many other possible worlds. Nowhere are there many sharpenings of the cat.

We can say that the sharpenings of a thing are just it but in another possible world.[24] Or else we can say that they are its counterparts there.[25] I shall not say more about the nature of sharpenings here because it does not matter, in the end, whether there really are any. In the following section, I shall introduce a class of abstract objects, the fuzzy boundaries of vague cats, countries, and other things. I shall also describe a procedure for sharpening up these boundaries. What really matters is that we can evaluate claims about how things would be, were they to have boundaries that are sharper than their actual boundaries. If the correct treatment of these counterfactuals does not call for possible objects, well and good. Then sharpenings will be suggestive fictions with a purely heuristic role.

### V. FUZZY BOUNDARIES

Since the whisker is coming loose it is incorrect to assert:

> The whisker is a part of Tibbles's head.

It is also incorrect to assert:

> The whisker is a part of Tibbles.

Suitable "many valued" truth tables can account for this, by allowing these sentences to go with truth-value gaps, neither true nor false.[26] But truth tables cannot account for the fact that it *is* correct to assert:

> If the whisker is a part of Tibbles's head, then it is a part of Tibbles.

---

[24] Compare Saul Kripke, *Naming and Necessity* (Cambridge: Harvard, 1980).

[25] Compare Lewis, "Counterpart Theory and Quantified Modal Logic," this JOURNAL, LXV, 5 (March 7, 1968): 113-26.

[26] See Steven Blamey, "Partial Logic," in D. Gabbay and F. Guenthner, eds., *Handbook of Philosophical Logic*, Volume 3 (Boston: Reidel, 1986), pp. 1-70.

since there is no truth-functional difference between this sentence and another, gotten by substituting other constituents with truth-value gaps. Acceptable truth tables do not call it true because they do not call all of them true. So what makes this sentence assertible? If not truth, what does it have that its constituents and all of the unassertible substitution instances lack?

It has *super truth*, truth no matter how completely sharp boundaries are drawn for everything, whisker, head, and cat. In this section, I shall introduce a class of abstract objects: fuzzy boundaries, some sharper than others. At the end of this section, I shall use these boundaries and their sharpenings to illustrate the supervaluation of claims about vague objects.[27] In the final section, I shall use them to construct an abstract domain of vague objects and to show that there does not have to be anything peculiar about the existence or identities of vague objects.

We can think of a *sharp* boundary as the vanishingly thin outline of a precise object, perhaps as a suitable set of points.[28] And we can think of a *fuzzy* boundary as the more or less thick outline of a vague object, within which a range of vanishingly thin outlines can be drawn. That is, we can think of it as a *constraint* on the drawing of sharp boundaries. To begin with, we can make this idea precise by thinking of a fuzzy boundary as the set of those sharp boundaries that it allows to be drawn. To sharpen a fuzzy boundary is then just to choose one of its elements.[29]

This conception is familiar and good, but it is in two ways too simple. Firstly, fuzzy boundaries so conceived do not reflect the higher-order vagueness of objects. A suitable elaboration is well known, though. We can let fuzzy boundaries be *sets* of boundaries, sharp or fuzzy. Then the order of an object's vagueness will be reflected in the degree of nesting of its boundary.[30]

---

[27] Henryk Mehlberg first made such a proposal in connection with vague singular terms in *The Reach of Science* (Toronto: University Press, 1958). Bas C. van Fraassen coined the term 'supervaluation' in "Singular Terms, Truth-Value Gaps, and Free Logic," this JOURNAL, LXIII, 17 (September 15, 1966): 481-95. Fine's (*op. cit.*) is the most sophisticated supervaluational treatment of linguistic vagueness.

[28] Compare Richard Cartwright, "Scattered Objects," in his *Philosophical Essays* (Cambridge: MIT, 1987), pp. 171-86.

[29] These boundaries are abstract mathematical entities, sets. Some think that the boundaries of concrete objects are also concrete. One might develop a notion of concrete vague boundaries using a suitably elaborated vague mereology. I shall not try here.

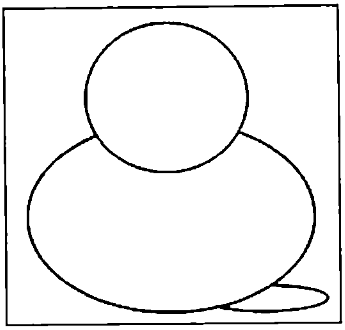
[30] Compare Fine's notion of an "n-th order boundary" (*op. cit.*, p. 293).

A second way in which this conception is too simple is that it does not require fuzzy boundaries to reflect the part-whole structure of objects. They ought to reflect it so that we can ensure that sharpening will satisfy a penumbral constraint. The cat's head, like the whole cat, is a vague object, having the loose whisker as a questionable part. Now, we cannot go calling the whisker a part of the cat's head unless we also call it a part of the whole cat; and we cannot go calling it a part of the cat unless we also call it a part of its head. So far, though, there is nothing to prevent this. We are quite free to choose from the boundary of the head an alternative that includes the whisker, but from the boundary of the cat an alternative that excludes it. We shall now further elaborate boundaries by adding *internal* boundaries, the boundaries of parts. Later this will enable us to ensure compliance with the penumbral constraint by sharpening fuzzy boundaries from the outside inwards. We shall sharpen the boundary of the cat by sharpening the boundaries of its head and other parts.

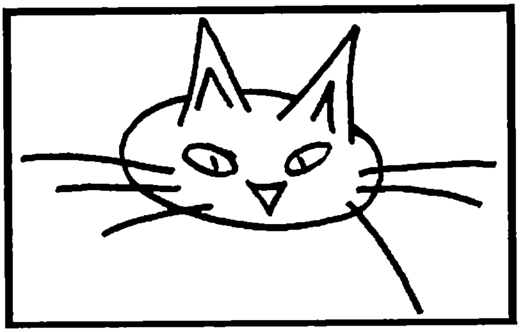
Let us begin with a stock of simple boundaries. These are to be the boundaries of mereological atoms. Taking simple boundaries as our basis, we inductively define successive levels of fuzzy boundaries. At each successor level there are two sorts of complex boundaries. There are nonempty sets of boundaries of lower levels, marked as *sharp collections*. These will be the boundaries of immediately precise objects. And there are other nonempty sets of boundaries of lower levels, marked as *vague collections*. These will be the boundaries of immediately vague objects.

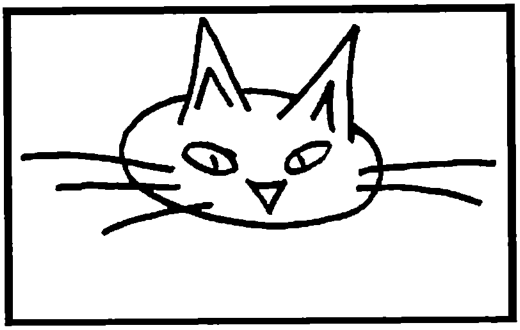
What is it for an object to *have* a boundary? Mereological atoms have basic boundaries, and only they have them. An object has a given sharp collection as its boundary if and only if this object is immediately precise, each of its immediate parts has a boundary within this collection, and each boundary within this collection is the boundary of one of its immediate parts. And an object has a given vague collection as its boundary if and only if this object is immediately vague, each of its immediate sharpenings has a boundary within this collection, and each boundary within this collection is the boundary of one of its immediate sharpenings.

Suppose Tibbles is immediately precise and, simplifying for the sake of the illustration, that her immediate parts are just head, torso, and tail. Suppose also that each of these immediate parts is immediately vague, having whiskers or hairs as its questionable immediate parts. Then, picturing sharp collections as rectangles and vague collections as ellipses, Tibbles's boundary, as far as the outer two levels of structure are concerned, looks like this:

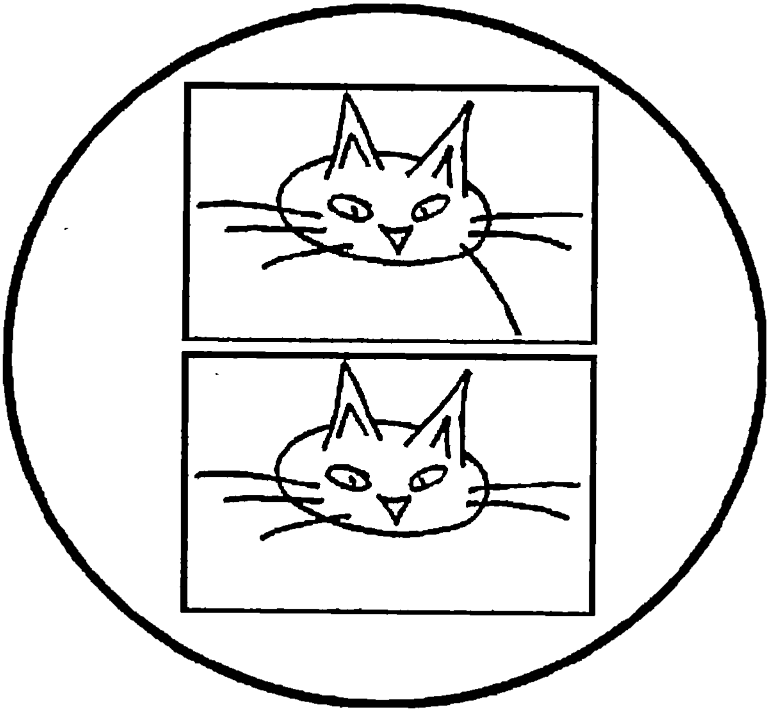


Continuing to oversimplify, suppose her head has just a single questionable immediate part, the whisker. Then it has just two immediate sharpenings, with and without. Second, let us declare the other parts of Tibbles's head (two ears, another five whiskers and so on) to be honorary mereological atoms. The boundaries of the two slight sharpenings of the head look like this:



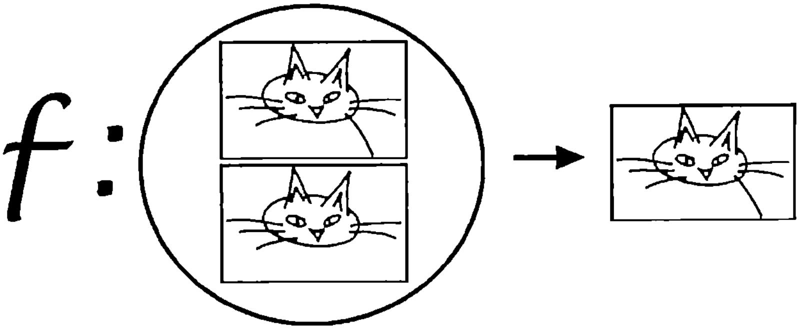


Finally, the boundary of Tibbles's head is the following vague collection:



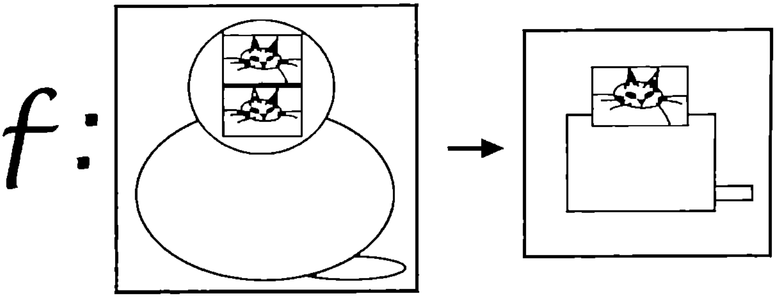
Now let us turn to the "drawing" of sharper boundaries. I shall describe incremental sharpening functions that, when applied repeatedly to fuzzy boundaries, yield *completely sharp* boundaries, within which there are no vague collections. Informally, the idea is to work from the outside of boundaries inwards, as we go replacing each vague collection with our choice from among its elements. Choices are to be consistent: we shall always choose the same element from any given vague collection, whenever we run up against it. When we run up against a sharp collection we pass straight to its elements, since to sharpen up an immediately precise object is just to sharpen up its immediate parts and then to put the results back together as they were. When we reach a basic boundary, we stop. It is the boundary of an atom and has no vagueness due to indeterminacy in part-whole relations.

Technically, a way of sharpening boundaries is a *choice function* $f$ that maps vague collections onto their elements: for each vague collection $V$, $f(V) \in V$. For example, the pictured choice function $f$ maps the boundary of Tibbles's head onto the boundary of the sharpening of which the loose whisker is a part.



Another choice function, say, $g$, maps this boundary onto the other alternative, without the whisker. To extend choice functions to arbitrary boundaries we put $f(B) = B$ if $B$ is a basic boundary, and $f(S) = [f(s){:}s \in S]$ if $S$ is a sharp collection (here '[' and ']', like the rectangles of the figures, enclose sharp collections).

Finally, $f$ maps Tibbles's boundary onto a precise boundary within which the boundary of the whisker is embedded.



Here, oversimplifying, the result of a single application of $f$ to the boundary of Tibbles's head is completely sharp. In general, it will take longer to achieve complete sharpness but we shall always get there in the end, since each application of $f$ reduces the degree of nesting of vague collections. Finite iteration maps any boundary $B$ onto a completely sharp boundary $f^*(B)$, the *complete sharpening* of $B$ by $f$.[31]

Using choice functions we can incrementally sharpen up the boundaries of vague objects. We have supposed that in other possible worlds there are objects, immediate sharpenings, which have the incrementally sharper boundaries. Finite iteration leads to completely sharp boundaries and *complete sharpenings*. Now we shall say that a claim is super true of some objects if it is true no matter how we sharpen them all up. Letting super truth of claims about completely sharp objects be truth according to some antecedently given evaluation scheme, super truth passes from claims about completely sharp objects to claims about vague objects of ever increasing orders of vagueness, one order at a time:

---

[31] For any sharp collections $S_i$, let $\cup_i S_i$ be the sharp collection [$s$: for some $i$, $s \in S_i$]. Now for any choice function $f$, $f(\cup_i S_i) = \cup_i f(S_i)$. Provided $\cup_i S_i$ is a boundary (the $S_i$ must be bounded in their complexity, say finite in number) $f^*(\cup_i S_i) = \cup_i f^*(S_i)$. We shall need this fact in the next section.

> Super-truth condition: a statement about some vague objects is *super true* (*super false*) if and only if, for every choice function, it is super true (super false) of their immediate sharpenings by that function.

For instance:

> The whisker is a part of Tibbles's head.

is neither super true nor super false. The complete sharpening of the whisker by the earlier choice function *f* is a part of the complete sharpening of Tibbles's head by *f*, but the same does not hold for the complete sharpenings by *g*. For the same reason:

> The whisker is a part of Tibbles.

has a super-truth gap.[32] But the "penumbral truth":

> If the whisker is a part of Tibbles's head, then it is a part of Tibbles.

is super true. This is just because any sharpening that like *f* counts the whisker a part of the head must also count it a part of the cat, since the head is a part, indeed an immediate part, of the cat. Choice functions respect part-whole structure.

Fuzzy sets, having their elements to a greater or lesser degree, are, like my fuzzy boundaries, a kind of vague abstract object.[33] Some think they can contribute to an understanding of the composition and identity of vague objects.[34] But there is a familiar and crucial difficulty with the use of fuzzy sets to explain vague composition.

---

[32] I assume that 'a part of' expresses a relation among completely precise objects that is at least as inclusive as the ancestral of the immediate part-whole relation. Any immediate part of...an immediate part of an object is to be a part of it.

[33] Lofti Zadeh, "Fuzzy Sets," *Information and Control*, VIII (1965): 338-53.

[34] van Inwagen uses fuzzy sets to explain vague composition (*op. cit.*, pp. 221-24). Jack Copeland uses them to explain vague identity in "Vague Identity and Fuzzy Logic," this JOURNAL, XCIV, 10 (October 1997): 514-34.

K.P.-1374

Standard many-valued interpretations of fuzzy logic are truth functional.[35] Such a treatment runs afoul of penumbral truth.

What of the vagueness of singular terms? Supposing neither Tibbles nor her head is vague, but instead the terms 'Tibbles' and 'Tibbles's head' are, we can still say that the penumbral sentence is super true. Now we shall mean by this true on all ways of resolving the reference of these terms. But we can say so only if we can somehow coordinate the resolution of singular terms which, like these two, pick out objects related as parts to wholes (or which come to do so, as a result of resolution). The idea is familiar in outline, but the details have still to be filled in.[36] One cannot simply mimic this treatment since, unlike these two, singular terms do not in general reflect the part-whole structure of things.

Taken at face value, supervaluation does something amazing: it reconciles vagueness with classical logic, finding truth-value gaps where, in view of the super truth of the law of the excluded middle, it might seem that there can be none. But maybe this is just a trick, worked by passing off as truth something that really is no such thing. Arguably, a disjunction of two sentences is true only if one of them is true; but sometimes a disjunction is super true although neither disjunct is super true. Truth is disquotational, but super truth is not.[37]

I think that super truth is not a sort of truth. Rather, it is a surrogate that is sometimes good enough: the assertibility of sentences goes by their truth values if they have them, but otherwise by their super-truth values. Indeed, we simply cannot identify truth with super truth—not if we insist that identities are always definite. In an interestingly full vague world, there are sure to be things that have some but not all of their complete sharpenings in common. Sentences expressing the identity of any two such objects will be true on some of their complete sharpenings but false on others. They will have super-truth gaps. If we say that super truth is truth, then we will also have to say that these identity sentences have truth-value gaps, not because their terms fail to pick anything out, but because of vagueness in what they pick out. We shall have to say that there are indefinite identities.

This completes the discussion of fuzzy boundaries. In the following section, I shall use them to construct an abstract vague world. We shall see that there is nothing especially mysterious about the idea of vague objects. In particular, it does not entail that composition and identities are vague. In fact, we shall see, this idea is even quite conservative, since it is compatible with the classical theory of parts and wholes.

---

[35] Joseph Goguen pioneered them in "The Logic of Inexact Concepts," *Synthese*, XIX (1969): 325-73.

[36] See Hartry Field, "Quine and the Correspondence Theory," *The Philosophical Review*, LXXXIII (1974): 200-28; and see Fine's notion of "penumbral connection" (*op. cit.*).

[37] Compare Timothy Williamson, *Vagueness* (New York: Routledge, 1994), p. 162.

## VI VAGUE MEREOLOGY

What is the part-whole theory of vague objects? I shall take the classical theory as my starting point. Letting 'part of' be primitive, one thing *overlaps* with another if they have some part in common. One thing is a *fusion* of some others if any further thing overlaps with it, the fusion, if and only if this further thing overlaps with one or other of the things fused. A fusion is pieced together from the things fused. Three basic propositions of classical mereology are:

Transitivity: the parts of the parts of anything are parts of it.
Unrestricted composition: any things have a fusion.
Unique composition: any things have a unique fusion.

A *proper part* of a thing is any part except the whole thing. Mereological atoms are not identifiable from their proper parts since they have none; but, according to another basic proposition of classical mereology, all other things are identifiable from their proper parts:

> Extensionality: if any given things have proper parts, and they have the same proper parts, then these things are identical.

Classical mereology is defined by these principles together with reflexivity, which says that anything is a part of itself, and antisymmetry, which says that no two objects are each other's parts. That is, classical part-whole relations are partial orders. In the presence of extensionality, unique composition is redundant.[38]

Classical mereology does not say much about things. It does not say whether they have temporal as well as spatial parts. Nor does it say whether they are made up of atoms or can be divided indefinitely. But it might be thought that classical mereology says that things are precise. Certainly, it has been thought (as I noted in section II) that, if things have questionable parts, then composition must be vague and thus restricted. It might also seem that, if things have questionable parts, there ought to be two things with the same proper parts but different questionable parts, refuting extensionality. In this section, I shall show that these basic propositions of classical mereology do *not* say that the world is precise.

---

[38] For a full introduction to classical mereology, see Peter Simons, *Parts: A Study in Ontology* (New York: Oxford, 1987)

Earlier, I argued that organisms are vague because they have temporary functional parts. It might seem a bit odd to argue now that vague objects can have a classical mereology because, unlike the relation of inclusion among quantities, functional part-whole relations are not an intended interpretation of classical mereology. Nor are some of these relations a plausible interpretation. If 'a part of' expresses some sort of direct functional part-whole relation, for instance, then there are counterexamples to transitivity: whereas its nucleus is a direct part of a cell, and the cell in turn a direct part of an organ, the nucleus is not a direct part of the organ.[39] I do not mean to suggest that mereology *has* to be classical, though. My point is just that it *can* be classical, even if there are vague objects. There might be good reasons to reject classical mereology. But the need to accommodate ontological vagueness, by itself, is not one of them.

I shall proceed by validating the propositions of mereology within an abstract domain of vague objects. Two ideas underlie its construction. The first is to take fuzzy boundaries themselves as objects—as vague objects that are their own boundaries. The second is to generalize the part-whole and fusion relations from precise to vague objects using a sort of ontological pendant of supervaluation. One thing—precise or vague—will be a (generalized) part of another if it is a part no matter how both are made completely sharp. A (generalized) fusion of some things will be something that is pieced together from them no matter how everything in question is made completely sharp. Proper definitions will follow. Meanwhile, notice that I shall not spell out generalized fusion, by analogy with the classical notion, in terms of the generalized part-whole relation. Instead, I shall spell out the generalized notions in terms of sharpenings and the corresponding classical notions.

Accordingly, we now need some precise objects with a classical fusion operation, some vague objects, and sharpening functions that map the vague onto the precise objects. First, the:

> *Precise objects*: these are nonempty sharp collections of simple boundaries. The part-whole relation among them is set-theoretic inclusion ($\subseteq$) and fusion is union ($\cup$).

Clearly, the part-whole theory of these objects is classical. In defining the vague objects, we can run together all fuzzy boundaries that, no matter how they are made completely sharp, turn out to be of the same simple stuff. To compare the simple stuff of *sharp* boundaries,

---

[39] Compare Nicholas Rescher, "Axioms for the Part Relation," *Philosophical Studies*, VI (1955): 8-11.

(nested) sharp collections of simple boundaries, we shall use an homogenizing function $H$. This is a function that maps sharp boundaries onto precise objects, in the above sense, by adding or removing structure as needed. For any simple boundary $B$, we put $H(B) = [B]$; and for any sharp boundary $S$, we put $H(S) = \cup_{s \in S} H(s)$. We say that boundaries $B$ and $C$ are *equivalent* if and only if for every choice function $f$: $H(f^*(B)) = H(f^*(C))$. $B^\sim$ is the class of all boundaries equivalent to $B$. Now we have the:

> *Vague objects*: These are equivalence classes $S^\sim$ of sharp collections, $S$.

For instance, letting $A$ and $B$ be basic boundaries, the vague collection $\langle B, [A,B] \rangle$ is a fuzzy boundary. (Square and angular brackets, like the rectangles and ellipses of the figures, respectively mark sharp and vague collections.) The equivalence class $[\langle B, [A,B] \rangle]^\sim$ of the sharp collection $[\langle B, [A,B] \rangle]$ is a vague object, in the sense of this definition.

It remains first to define sharpening functions that map vague onto precise objects, and then to extend the relations of part-whole and fusion from precise to vague objects. Vague objects will take their sharpenings from their representatives: for any choice function $f$ we let $f^*(S^\sim)$, the *complete sharpening* of $S^\sim$ by $f$, be $H(f^*(S))$. Notice that $f^*(S^\sim)$ is always a precise object, in the sense of this construction. We generalize the part-whole relation by maximizing definite parthood and nonparthood. Thus $S^\sim$ is a (*definite*) *part* of $T^\sim$ if and only if for each $f$: $f^*(S^\sim) \subseteq f^*(T^\sim)$, and $S^\sim$ is a (*definite*) *nonpart* of $T^\sim$ if and only if for each $f$: $f^*(S^\sim) \not\subseteq f^*(T^\sim)$. If neither a part nor a nonpart, then $S^\sim$ is a *questionable part* of $T^\sim$. Clearly, these objects really do have questionable parts. For example, $[A]^\sim$ is a questionable part of $[\langle B, [A,B] \rangle]^\sim$. Letting $f$ and $g$ be such that $f(\langle B, [A,B] \rangle) = [A,B]$ but $g(\langle B, [A,B] \rangle) = B$, we see that $f^*([A]^\sim) = [A] \subseteq [A,B] = f^*([\langle B, [A,B] \rangle]^\sim)$, but $g^*([A]^\sim) = [A] \not\subseteq [B] = g^*([\langle B, [A,B] \rangle]^\sim)$.

Although these objects are vague, there is nothing shady about their presence. They are fully there. There is nothing vague about their identities, either: these objects are identical to themselves, they are different from other things, and they are indefinitely identical to nothing. They refute Evans's argument from definite identities that there cannot be vague objects.

Let us say that an object is *utterly* vague if there is no sharp line between *any* of its penumbral categories: not between its parts and its nonparts, not between its parts and its questionable parts, nowhere. The vague objects of this construction have sharp sets of parts, questionable parts, and nonparts, and are not utterly vague. I expect this construction can be generalized to admit higher-order vagueness (I have not shown that it can) but the point would remain. Objects would still have sharp sets of absolutely uncontroversial parts and nonparts—of things that are in no sort of limbo at all.[40] These vague objects suit my purposes quite well, although they are perhaps not quite as vague as can be, because they can be constructed with simple and familiar means.

Let us now verify the four basic propositions of mereology. To start with, there is:

> *Transitivity*: if $S^{\sim}$ is a part of $T^{\sim}$ and $T^{\sim}$ is a part of $U^{\sim}$, then $S^{\sim}$ is a part of $U^{\sim}$.

I assume it is sufficient to show that any three objects that verify the antecedent must also verify the consequent. Suppose $S^{\sim}$ is a part of $T^{\sim}$, and $T^{\sim}$ is a part of $U^{\sim}$. Take any choice function $f$. Then $f^{*}(S^{\sim}) \subseteq f^{*}(T^{\sim}) \subseteq f^{*}(U^{\sim})$ so, by transitivity of $\subseteq$ and arbitrary choice of $f$, $S^{\sim}$ is a part of $U^{\sim}$.

Since we must reckon with truth-value gaps, transitivity might be thought to require in addition that any three objects that falsify the consequent of an instantiation will also falsify its antecedent. But transitivity fails on such a double-barreled interpretation of 'if..., then...'. For instance, letting $A$, $B$, $C$, and $D$ be different atoms, choose $S = [\langle B,[A,B]\rangle]$, $T = [\langle C,[B,C]\rangle]$, and $U = [\langle D,[C,D]\rangle]$. Then $S^{\sim}$ is a nonpart of $U^{\sim}$, so these three objects falsify the consequent of the relevant instantiation. But it is neither the case that $S^{\sim}$ is a nonpart of $T^{\sim}$, nor the case that $T^{\sim}$ is a nonpart of $U^{\sim}$. To falsify a conjunction is to falsify a conjunct, so these objects fail to falsify the antecedent.

Another basic fact is that no two objects are each other's parts: if $S^{\sim}$ is a part of $T^{\sim}$ and $T^{\sim}$ is a part of $S^{\sim}$, then $S^{\sim}=T^{\sim}$. It is straightforward to show that whenever the antecedent of an instantiation of this claim is satisfied, so is its consequent. Once again, sometimes the consequent of an instantiation is false but not its antecedent: $[\langle A,B\rangle,\langle C,D\rangle]^{\sim}$ and $[\langle A,D\rangle,\langle C,B\rangle]^{\sim}$ are two objects, and each is a questionable part of the other.

Now let us generalize the fusion relation and verify the remaining basic propositions of mereology. We say that $F$ is a (*definite*) (*generalized*) *fusion* of some objects $S_1^{\sim}$, $S_2^{\sim}$, $S_3^{\sim}$,...if and only if for all $f$: $f^{*}(F)$ is a classical fusion of $f^{*}(S_1^{\sim})$, $f^{*}(S_2^{\sim})$, $f^{*}(S_3^{\sim})$,.... We say that $F$ is a (*definite*) (*generalized*) *nonfusion* of $S_1^{\sim}$, $S_2^{\sim}$, $S_3^{\sim}$,...if and only if for all $f$: $f^{*}(F)$ is not a classical fusion of $f^{*}(S_1^{\sim})$, $f^{*}(S_2^{\sim})$, $f^{*}(S_3^{\sim})$,.... Otherwise, $F$ is a *questionable* (*generalized*) *fusion* of $S_1^{\sim}$, $S_2^{\sim}$, $S_3^{\sim}$,....

---

[40] For reasons that Mark Sainsbury sets out (*op. cit.*, p. 255).

To illustrate, here is a version of the informal example of section II, which suggests that questionable parthood need not restrict composition. Let $W$ be one boundary, and let $C'$ be another. Let Wisp be $[W]^{\sim}$, let Cloud-Minor be $[C']^{\sim}$, let Cloud be $[\langle C',[C',W]\rangle]^{\sim}$, and let Cloud-Major be $[C',W]^{\sim}$. Then Wisp is a questionable part of Cloud, and Cloud is a questionable fusion of Wisp and Cloud-Minor. Even so Wisp and Cloud-Minor compose to make up something: Cloud-Major is a fusion of them.

Let us for the moment restrict attention to vague objects with representatives whose complexity lies below some given finite upper bound. Then we do indeed have:

> *Unrestricted Composition*: any vague objects $S_i^{\sim}$ have a fusion, namely, $(\cup_i S_i)^{\sim}$.

This follows directly from the fact that for any $f$, $f^*((\cup_i S_i)^{\sim}) = \cup_i f^*(S_i^{\sim})$. This fact follows, in turn, from the earlier observation that for any $f$ and sharp collections $S_i$, $f^*(\cup_i S_i) = \cup_i f^*(S_i)$, together with the easily verified further observation that $H(\cup_i S_i) = \cup_i H(S_i)$. The restriction to $S_i^{\sim}$ such that the $S_i$ are of limited complexity ensures that $\cup_i S_i$ is also a suitable boundary, and that $(\cup_i S_i)^{\sim}$ is an object. Without this restriction, we cannot always be sure that $\cup_i S_i$ is even a boundary; it will be if the $S_i$ are finite in number, though, so even without this restriction there are always finite fusions.

The third basic proposition of classical mereology is:

> *Unique composition*: if $F^{\sim}$ and $G^{\sim}$ are generalized fusions of the $S_i^{\sim}$, then $F^{\sim} = G^{\sim}$.

This can be seen as follows. Letting $F^{\sim}$ and $G^{\sim}$ be generalized fusions of the $S_i^{\sim}$, it follows from the properties of classical fusions that for any $f$, $f^*(F^{\sim}) = f^*(G^{\sim})$, since whatever overlaps with either also overlaps with the other. So for every $f$: $H(f^*(F)) = f^*(F^{\sim}) = f^*(G^{\sim}) = H(f^*(G)$, and so $F^{\sim} = G^{\sim}$.

Finally, let us see that the vague domain is extensional. It is sufficient to show that everything conforms to a:

> *Principle of proper parts*: if some vague object has proper parts, and each of its proper parts is a part of some second object, then the object in question is also a part of this second object.

Extensionality follows immediately since no two objects are each other's parts. We can demonstrate this principle within the full domain of vague objects, with representatives of arbitrary complexity.[41] The proper parts that make a difference between any two vague objects with proper parts are provided by the following lemma:

> *Lemma*: if $S^{\sim}$ is a proper part of $T^{\sim}$, then $[\langle S,T\rangle]^{\sim}$ is also proper part of $T^{\sim}$.

The demonstration of this lemma is as follows. Let $V$ be the vague collection $\langle S,T\rangle$. Then for any $f$, either $f^*(V) = f^*(S)$, or $f^*(V) = f^*(T)$. That $[V]^{\sim}$ is a part of $T^{\sim}$ is clear: for any $f$, $f^*([V]^{\sim}) = H(f^*([V])) = H([f^*(V)]) =$ either $H(f^*(S))$ or $H(f^*(T)) =$ either $f^*(S^{\sim})$ or $f^*(T^{\sim})$. Either way, since $S^{\sim}$ is a part of $T^{\sim}$, $f^*([V]^{\sim}) \subseteq f^*(T^{\sim})$. To see that $[V]^{\sim} \neq T^{\sim}$, since $S^{\sim} \neq T^{\sim}$ we can choose $f$ such $H(f^*(S)) \neq H(f^*(T))$. Let $f_{[V/S]}$ be the choice function that differs from $f$, if at all, only insofar as $f_{[V/S]}(V) = S$. Since by the well-foundedness of boundaries $V$ appears neither within $S$ nor within $T$, we now have: $H(f_{[V/S]}{}^*([V])) = H([f_{[V/S]}{}^*(S)] = H[f^*(S)] = H(f^*(S)) \neq H(f^*(T)) = H(f_{[V/S]}{}^*(T))$. So $[V]^{\sim} \neq T^{\sim}$. This completes the demonstration of the lemma.

The principle of proper parts follows as a consequence. Supposing it is not the case that $T^{\sim}$ is a part of $U^{\sim}$, choose $f$ such that $H(f^*(T)) \not\subseteq H(f^*(U))$. Supposing $T^{\sim}$ has a proper part, since $U$ has a finite degree of nesting it is possible, with repeated use of the lemma, to find a proper part $[V]^{\sim}$ of $T^{\sim}$ such that $T$ is an element of the vague collection $V$, but $V$ does not occur within $U$ (nor, by the well-foundedness of boundaries, does it occur within $T$). Now: $f_{[V/T]}{}^*([V]^{\sim}) = H(f_{[V/T]}{}^*[V]) = H[f_{[V/T]}{}^*(V)] = H(f_{[V/T]}{}^*(V)) = H(f_{[V/T]}{}^*(T)) = H(f^*(T)) \not\subseteq H(f^*(U)) = H(f_{[V/T]}{}^*(U)) = f_{[V/T]}{}^*(U^{\sim})$. So $[V]^{\sim}$, though a proper part of $T^{\sim}$, is not a part of $U^{\sim}$. This verifies the principle of proper parts, and with it extensionality.

It might appear that extensionality is at odds with the earlier idea that vague material objects are three-dimensional continuants, constituted by precise quantities of matter but not identical to them. Do Tibbles and the matter that constitutes her not have the same proper parts? Does extensionality not require that they are identical? It does not. If Tibbles is vague, then there is no unique quantity of matter that constitutes her. There are many quantities, with equal claims to constitute her, and none has the same proper parts as Tibbles. Consider first any candidate that is a part of Tibbles. It is a proper part of her (since they are not identical), but not a proper part of itself. Now consider any candidate that is not a part of Tibbles. Being precise, it is its own complete sharpening. By the definition of generalized parthood, there is some complete sharpening of Tibbles of which this candidate is not a part. Choose any proper part of the candidate that is not a part of the sharpening in question. (The extensional mereology of quantities provides one if we assume that complete sharpenings, like constitution candidates, are quantities of matter.) By the definition of generalized parthood, the chosen proper part of the candidate is not a part of Tibbles, and therefore not a proper part of her.

---

[41] As I have noted, we can now only be sure to have finite fusions.

We have seen that there is nothing especially mysterious about the idea of worldly vagueness. Vague objects need not have any sort of shady presence. They can be fully there. Nor need they have dodgy identities. They can be identical to themselves, different from other things, and indefinitely identical to nothing. And there is an end of it.

MICHAEL MORREAU

University of Maryland/College Park

"how to think in new ways"*

DELEUZE
PARNET

**After Christianity**
**Gianni Vattimo**
**Translated by**
**Luca D'Isanto**

Taking up the challenge presented by Nietzche's pronouncemnet of the "death of God," one of Europe's foremost philosophers contemplates the future of religion in the postmodern world.

$24 50 cloth
ITALIAN ACADEMY LECTURES

NOW IN PAPER
**Dialogues**
Second Edition
**Gilles Deleuze and Claire Parnet**
**Translated by**
**Hugh Tomlinson**

"The best introduction to Deleuzian philosophy. A dazzling exposition of Deleuze's concepts and methodologies."
—Eric Alliez*

$19 50

Sign up for bulletins of new books in your field at
www.columbia.edu/cu/cup/catalog

**COLUMBIA UNIVERSITY PRESS**
columbia.edu/cu/cup 800-944-8648

# MATHEMATICS AND METAPHILOSOPHY*

The philosophy of mathematics is philosophy in a pure state, stripped of all worldly appendages; austere; philosophy without the sugar coating of a pretense to Relevance to Life; but also heady, perhaps best taken in short sips.

—Paul Benacerraf[1]

Mathematical realism and traditional rationalism have been moribund for many decades. To be sure, there have been recent defenses of each, but they have not changed the overall situation significantly. The reason is not, as many philosophers suppose, that the criticisms of mathematical realism and traditional rationalism are so good that attempting to defend them is pointless. The problem is not flawless criticisms, but flawed defenses.

What is the flaw which has proven fatal to the defenses of mathematical realism and traditional rationalism, and which, if I am right about the criticisms of these positions, is largely responsible for their unfortunate condition? I shall argue that it is separatism: realism is defended separately from rationalism and rationalism separately from realism. As a consequence, realism is epistemologically vulnerable and rationalism metaphysically incomplete. Given this diagnosis, the treatment plan is obvious: integrate realism and rationalism. The prognosis is that an integrated position contains the resources, on the one hand, to provide adequate replies to antirealist criticisms like Paul Benacerraf's and antirationalist criticisms like W. V. Quine's and, on the other, to complete the metaphysical basis for a rationalist conception of mathematical knowledge.

Implementation of this plan began with my *Realistic Rationalism.*[2] Here, I shall significantly extend the earlier account of how knowledge of abstract mathematical objects is possible, and provide a new version of the reply to Benacerraf's[3] epistemic challenge to mathe-

---

* Thanks for helpful comments to Jody Azzouni, Bradley Armour-Garb, Charles Chihara, Jonathan Kastin, Elliott Mendelson, David Pitt, and my philosophy audiences at the University of Nebraska, University of British Columbia, the University of Delaware, and the CUNY Graduate Center, where earlier versions of this paper were presented.

[1] "What Mathematical Truth Could Not Be," in Adam Morton and Stephen P. Stich, eds., *Benacerraf and His Critics* (Cambridge: Blackwell, 1996), pp. 9-59, here p. 10.

[2] Cambridge: MIT, 1998.

[3] "Mathematical Truth," this JOURNAL, LXX, 19 (November 8, 1973): 661-79; reprinted in Benacerraf and Hilary Putnam, eds., *The Philosophy of Mathematics:*

0022-362X/02/9907/362–90

matical realism. I shall also answer objections that have been made to the reply, and, most importantly, I shall add a vindication of mathematical intuition against charges ranging from unreliability to scandalous mysticism, one which shows that mathematical intuition is not only innocent but indispensable to any adequate account of our epistemic faculties. Further, I shall argue (a) that the principal defenses of mathematical realism and traditional rationalism—Kurt Gödel's, Michael Resnik's, and Stewart Shapiro's defenses of the former, and George Bealer's and Thomas Nagel's defenses of the latter—fail because of their separatism and (b) that a unified realism/rationalism succeeds because of its integration.[4] Finally, I shall explain how a defensible mathematical realism provides the basis for a philosophical ascent to a metaphilosophy that offers an alternative to Ludwig Wittgenstein's linguistic positivism and Quine's holistic empiricism.

The argument here begins with the failure of Gödel's mathematical realism to meet Benacerraf's epistemic challenge. For two reasons, I focus on Benacerraf's epistemic challenge in "Mathematical Truth" rather than his semantic challenge in "What Numbers Could Not Be."[5] First, Benacerraf's semantic challenge—his argument that reference to numbers is not reference to objects—does not target mathematical realism per se. Classical mathematical realism takes numbers to be objects, but, as mathematical structuralism makes clear, realism itself need not take them in this way. Second, Benacerraf's semantic challenge is the one on which antirealists have consistently relied in attacking realism. Dale Gottlieb[6] says: "*Abstract entities are mysterious and must be avoided at all costs*" (*ibid.*, p. 11); Hartry Field[7] says that the realist "is going to have to postulate some *aphysical*

---

*Selected Essays* (Englewood Cliffs, NJ: Prentice-Hall, 1964), pp. 403-20 (page references in this article are to the reprinted version).

[4] Gödel, "What Is Cantor's Continuum Problem?" *The American Mathematical Monthly*, LIV (1947): 515-25, reprinted in Benacerraf and Putnam, pp. 258-73, and "Some Basic Theorems on the Foundations of Mathematics and Their Implications," in *Kurt Gödel. Collected Works, Volume III*, Solomon Feferman et al., eds. (New York: Oxford, 1995), pp. 304-23; Resnik, *Mathematics as a Science of Patterns* (New York: Oxford, 1997); Shapiro, *Philosophy of Mathematics: Structure and Ontology* (New York: Oxford, 1997); Bealer, "*A Priori* Knowledge and the Scope of Philosophy," *Philosophical Studies*, LXXXI (1996): 121-42; and Nagel, *The Last Word* (New York: Oxford, 1997).

[5] *The Philosophical Review*, LXXIV, 1 (January 1965): 47-73; reprinted in Benacerraf and Putnam, pp. 272-94.

[6] *Ontological Economy: Substitutional Quantification and Mathematics* (New York: Oxford, 1980).

[7] *Science without Numbers* (Princeton: University Press, 1980); see also his *Realism, Mathematics, and Modality* (New York: Blackwell, 1989).

*connection*, some *mysterious grasping*" (*ibid.*, p. 59); Charles Chihara[8] says that the realist's appeal to intuition is "like appealing to experiences vaguely described as 'mystical experiences' to justify belief in the existence of God" (*ibid.*, p. 215); Michael Dummett[9] says intuition "has the ring of philosophical superstition" (*ibid.*, p. 202).

The assumption behind such rhetoric—that the impossibility of natural contact with abstract objects condemns mathematical realists to an epistemology of supernatural contact—holds only if there is no alternative to a contact epistemology or no epistemology at all. But there is a third possibility: a no-contact epistemology. Neither Benacerraf nor other antirealists have given an argument that rules it out. Moreover, the point that they have been at such pains to stress—the impossibility of causal contact with abstract objects—serves equally well as a basis on which to argue for a no-contact epistemology. For antirealists, this impossibility is a reductio of the notion that a realist ontology can have an adequate epistemology, but only because epistemology, for them, is contact epistemology. But, not having a commitment to empiricism, realists can turn this reductio on its head. They can say that the impossibility of basing mathematical knowledge on causal contact shows that mathematical knowledge must be based on something other than causal contact. One philosopher's modus ponens is another's modus tollens. Since, from the realist's standpoint, the antirealist's mistake is the equation of epistemology with contact epistemology, the way is now open for an explanation of mathematical knowledge which makes no appeal to contact, either natural or supernatural.

A cure for the moribundity of mathematical realism provides a cure for the moribundity of traditional rationalism. The moribundity of rationalism was due to the popularity of Wittgensteinian and Quinean naturalism. If all knowledge is fundamentally empirical knowledge of natural objects, then rationalism is either linguistic nonsense, as both the early and late Wittgenstein argued, or empirically false, as Quine argued. On the other hand, if knowledge in some domain of science is fundamentally about non-natural objects, then this domain provides a basis for an ascent to a rationalist metaphilosophy. Thus, saving mathematical realism from Benacerraf's epistemic criticism enables us to take mathematics as a domain of science in which knowledge is not natural knowledge, and, hence, to take mathematics as the basis for an ascent to a rationalist metaphilosophy. This ascent would be similar to the "semantic ascents" familiar from the work of Rudolf Carnap, Wittgenstein, and Quine, but, whereas their metaphilosophies strip philosophy of its standing as an a priori study of the most general facts about being and knowledge, the one I am advocating restores that standing to philosophy.

---

[8] "A Gödelian Thesis Regarding Mathematical Objects: Do They Exist? And Can We See Them?" *The Philosophical Review*, XCI (1982): 211-27.

[9] "Platonism," in *Truth and Other Enigmas* (Cambridge: Harvard, 1978), pp. 202-14.

## I. MATHEMATICAL KNOWLEDGE

*I.A. Mathematical knowledge: a reply to Benacerraf.* Benacerraf[10] argues as follows. If it is claimed that *X* knows a true proposition *p* about numbers, and if, as the classical realist claims, numbers are abstract objects, then, since such objects are atemporal, aspatial, and hence causally inaccessible to us, *X* can have no contact with them. If "*X*'s four-dimensional space-time worm does not make the necessary (causal) contact with the grounds of the truth of the proposition [*p*]," how then, asks Benacerraf, could *X* "have come into possession of the relevant evidence or reasons[?]" (*ibid.*, p. 413). Benacerraf continues:

> ...we use *p* to determine the range of possible relevant evidence. We use what we know of *X* (the putative knower) to determine whether there could have been an appropriate kind of interaction, whether *X*'s current belief that *p* is causally related in a suitable way with what is the case because *p* is true.... If [his evidence is not drawn from the range determined by *p*], then *X* could not know that *p*. The connection between what must be the case if *p* is true and the causes of *X*'s belief can vary widely (*ibid.*, pp. 413-14).

Benacerraf's argument purports to present mathematical realists with a dilemma: either adopt a contact epistemology and be stuck with the incoherent thesis that we have causal contact with causally inaccessible objects, or reject a contact epistemology and allow unconstrained variation that makes nonsense of the claim to have "relevant evidence or reasons" for our mathematical beliefs. In either case, the justification required for *X*'s belief in *p* to count as knowledge does not exist.

A no-contact epistemology would enable us to go between the horns of the dilemma. What is required is to provide a noncausal relation that establishes an unvarying connection between what must be the case if *p* is true and the causes of *X*'s belief in *p*. One reason for thinking there is such a relation is that something more than a causally based epistemology is required for an adequate account of mathematical knowledge. Even if the notion of grounding mathematical knowledge on contact with mathematical objects were not incoherent, such grounding could not provide an epistemology for mathematics. Supposing the mathematician to have the same relation to the actual as the natural scientist, the information obtained on the basis of the relation would not provide suitable grounds for mathematical knowledge because such knowledge is knowledge of necessary truth.[11] Experience, as Immanuel Kant pointed out, can teach us what *is* the case but not what *must be.*

---

[10] "Mathematical Truth," p. 413.

Causal contact has an epistemic point in natural science because natural objects can be otherwise than they actually are (non obstante their essential properties). Contact enables natural scientists to discover how natural objects actually are by providing empirical evidence for or against hypotheses about how they might possibly be. The epistemic function of causal contact is to determine which possibilities are actualities. Abstract objects like numbers cannot be otherwise than they actually are, at least not with respect to their mathematical properties. Since abstract objects of type *T* have their *T* properties necessarily, empirical methodology, exported from natural science to formal science, would get us nowhere.

An epistemology based on causal contact puts the wrong question in the mathematician's mouth. It represents the mathematician as asking the same question about mathematical objects as the natural scientist asks about natural objects, namely, (Q1):

> (Q1) Which of the rival possibilities of how such-and-such objects are is how they actually are?

But, as the above reflections indicate, the question the mathematician asks is rather (Q2).

---

[11] Saul Kripke—*Naming and Necessity* (Cambridge: Harvard, 1980), pp. 34-35—and Thomas Tymoczko—"The Four-Color Problem and Its Philosophical Significance," this JOURNAL, LXXVI, 2 (February 1979): 57-82—view the appeal to a computer computation appearing as a step in a proof as, in Tymoczko's way of putting it, "a report on a successful experiment" (p. 63). Viewed in this way, the appeal establishes the mathematical theorem "on grounds that are in part empirical." Hence, some mathematics is neither necessary nor a priori. I believe, however, that viewing the appeal to computer computation in this way is a conceptual confusion. Tymoczko himself comes close to recognizing this when at one point he wonders how "we can say that the computer registered a phenomenon of mathematical space" (p. 78). The confusion is generated by thinking of computer computation as an empirical experiment instead of as a priori reasoning in a nonhuman system. Chess-playing machines that chess champions use to prepare for tournaments are a better analogue than laboratory experiments. Such machines are parallel in all relevant respects to human coaches, which suggests that their solutions to chess problems, like those of human coaches, are a priori.

(Q2) Which of the rival suppositions about how such-and-such objects are is how they necessarily are?

Thus, the range of alternatives in inquiries in natural science and mathematics are different. The alternatives in natural science are possibilities: the hypothesis that something is a golden mountain may be farfetched, but it qualifies as an alternative, whereas the hypothesis that something is a round square does not. The alternatives in mathematics are supposabilities; for example, the hypothesis (in a reductio proof, say) of a largest number qualifies as a supposition, whereas the hypothesis of a blue number does not.

Furthermore, the outcomes of inquiries in natural science and mathematics are also different. The truths that natural sciences establish are contingent, while the truths that mathematics establishes are necessary. Hence, the notion that information about how objects actually are plays a role in mathematical justification misunderstands the fundamental difference between the formal and natural sciences.

Given that mathematics asks questions of the form (Q2), mathematical justification must be a matter of reason, largely of proving theorems. This assertion is at once a commonplace of mathematical practice and the thesis of traditional rationalism. Reason shows that one supposition is the truth about mathematical objects by exhibiting aspects of their structure which remove any possibility for rational doubt that they are otherwise than as that supposition represents them. The other suppositions about the mathematical object express the possibilities of rational doubt, so that excluding all of them excludes all rational doubt. Proofs thus establish the necessity of their conclusion because their reasoning is so tight that no room is left for rational doubt.

How can proofs establish the necessity of their conclusion when they talk about numbers, sets, and so on, and properties like being prime, being perfect, and so on, and typically make no mention of the modal status of the statements that comprise them? Here, realism supports rationalism. Given that mathematical objects are abstract objects, they have their mathematical properties necessarily. There is no such thing as an abstract object having a property that it might not have had. 6 being a perfect number, there is no coherent counterfactual in which 6 is not a perfect number. For statements about mathematical abstracta and their mathematical properties, truth collapses into necessary truth.

Quine[12] maintains both a robust mathematical realism on which "abstract entities...are the substance of mathematics" and an uncompromising empiricism on which all statements are contingent, revisable in principle on the basis of experience (*ibid.*, p. 45). This aspect of Quine's holistic empiricism should not be considered a counterexample to the claim that mathematical realism entails that proofs establish necessary truths. Rather, it is the other way around: this entailment is a counterexample to Quine's position. For, given its commitment to mathematical realism, the truth that 6 is a perfect number is equivalent to the necessary truth that 6 is a perfect number. Since mathematical truths are not on a par ontologically with truths about physical objects, they are not, as Quine (*ibid.*, p. 45) claims, on a par epistemologically with truths about them. As we shall see later, this is not the only problem for Quine's epistemology.

We are now in a position to complete a realist reply to Benacerraf's dilemma. The dilemma's weak spot is that the alternatives it poses are not exhaustive. They seem so because it is assumed that the formal sciences are epistemically like the natural sciences. The assumption is false: unlike natural science, which uses observation to prune down what is possible about concrete objects to what is actual, formal sciences use reason to prune down what is supposable about abstract objects to what is necessary. Relative to this difference, we can explain why $X$'s relation to what must be the case if $p$ is true does not have to be causal to restrict the unwanted variation between what must be the case if $p$ is true and the causes of $X$'s belief that $p$. If the cause of $X$'s belief in $p$ is reasoning of $X$'s that correctly and with proper understanding recapitulates a proof that $p$, then the cause of $X$'s belief precludes as impossible every supposition other than $p$ about mathematical reality. In this case, the relation between the cause of $X$'s belief and what must be the case if the belief is true, though not a causal relation, is, in virtue of the preclusion, not subject to variation.

I would now like to distinguish a reply to Benacerraf's epistemic challenge from an epistemology for mathematical realism. His epistemic challenge is, in a sense, a Kantian question: How is mathematical knowledge possible if, as realists claim, it is about objects with which we can have no causal contact? My reply explains how knowledge of the abstract is possible. It does not try to develop the explanation into a full-blown theory of mathematical knowledge. This would be a large undertaking, requiring, just to mention two things, canvassing the varieties of mathematical knowledge and situating the theory in current epistemology. There is, however, one important topic that may reasonably be considered as much a matter of reply as of theory, namely, knowledge of mathematical truths not subject to proof either because they are too basic or because they are too theoretical. I shall say more about this topic below.

---

[12] *From a Logical Point of View* (Cambridge: Harvard, 1953).

I also want to distinguish both a reply and an epistemology from an argument for mathematical realism. Realists and antirealists both have the philosophical obligation to justify their position, but their attempts to meet this obligation constitute the larger dialectic that began with Plato's realism and Aristotle's antirealism. Now, given that antirealists often use Benacerraf's epistemic challenge in making the case for their conception of mathematical reality, my reply to the challenge plays a role in my overall justification for mathematical realism; but the response has a limited, primarily defensive, role, one which, in the nature of things, cannot settle the issue at the heart of the dialectic.

I shall say a few more things about the argument for mathematical realism, but the topic goes beyond the scope of the present article. Nonetheless, I need to explain, at least briefly, what is and what is not required of such an argument.

In both the formal sciences and natural sciences, I picture knowledge as a hierarchical structure with those sciences themselves at the top level, and with philosophy, divided into the philosophical foundations of the sciences and the philosophical foundations of the foundations of the sciences, at the two successive levels below:

(H) First level: the sciences proper (mathematics, physics, and the like); the second level, the foundations of the sciences (philosophy of mathematics, philosophy of physics, and the like); the third level, the foundations of the foundations of the sciences (metaphysics).

At the first level, the arguments among theories are about the best explanation of the phenomena in their domain of science. At the second level, the arguments among theories are about the best explanation of philosophical aspects of the science in question. Typically, the issues, such as the ontological status of its objects, are not systematically addressed in the science itself. At the third level, the arguments are also among philosophical theories, but these address problems, especially skepticism, not addressed at the previous level. It is a curious, but nonetheless real, possibility that we might end up knowing that some second-level theory provides the best foundations for some science, but have to concede, from the perspective of the third level, that we have no basis for claiming that we know what the science says we do.[13]

The principal consequence I want to draw from this picture is that second-level theories are not open to either of the two forms of skepticism. Not being at the first level, mathematical realism is not open to unphilosophical skepticism of the type found in mathematics proper—for example, such as might arise from reflection on an unrestricted use of a comprehension principle for set formation or an unrestricted use of the concept of truth. Theories in the foundations of mathematics assume mathematical knowledge and try to show that they provide a better understanding of it than the rival theories. Not being at the third level, mathematical realism is not open to philosophical skepticism. Neither side in the controversy between realism and antirealism can use skepticism against the other because both sides are vulnerable. If a realist epistemology based on intuition and reason had to contend with René Descartes's skepticism about our knowledge of necessary truth, an antirealist epistemology based on perception and induction would have to contend with Descartes's and David Hume's skepticism about contingent truth.[14]

*I.B. Mathematical knowledge: elaborations.* I now want to elaborate on some of the ideas in my reply to Benacerraf's challenge in the process of responding to direct and indirect objections to it. We can proceed systematically. Since the central notion in my reply is that of proof, and since a proof is a finite sequence of propositions, each member of which is a premise or consequence of preceding propositions in the sequence, an objection will perforce target either what my reply says about premises or what it says about inferential steps. Since, furthermore, a premise can be either a proposition about the existence of mathematical objects or a proposition about the possession of mathematical properties, a possible objection can fall into one of three categories. Either it concerns what we say about existence premises, what we say about inferential steps, or what we say about possession premises.

*I.B.1. Mathematical knowledge: existence premises.* Jody Azzouni[15] raises a criticism in connection with existence premises:

---

[13] This is no endorsement of foundationalism. There is nothing here to suggest mathematics rests on philosophy. On my account, the absence of a philosophy of mathematics does not mean the absence of mathematics, only the absence of answers to certain philosophical questions about mathematics.

[14] For further discussion, see my *Realistic Rationalism*, pp. 51-58.

[15] "Stipulation, Logic, and Ontological Independence," *Philosophia Mathematica*, VIII (2000): 225-43.

> The traditional (and natural) objection to this sort of claim focuses on the premises used in a proof. Generally, it is such *premises* that provide the substantial ontological assumptions arising in mathematics. The axioms of first-order PA tell us, among other things, that there is a zero, and that every number has a successor. What in proofs, as Katz has described them, excludes the possibility of such premises being false (*ibid.*, p. 234)?
>
> [The problem is that] no argument has been given for *why* systematic considerations are dependable with respect to the ontologically independent mathematical objects that the theory is purportedly about (*ibid.*, pp. 234-35).

Azzouni objects that I neglected to argue inter alia that there is a number 0 and that there is no largest number, but that objection is off the mark. My reply and the epistemology to which it leads are in the foundations of mathematics, not in mathematics proper. It is not the philosopher's business to do the mathematicians' work for them. I have to take for granted their reasons for thinking there is a 0 but no largest number, since it is with such mathematical facts that my attempt to provide the best explanation begins.

If, instead, I construe Azzouni's objection as directed against my epistemology as a theory in the foundations of mathematics, it would be that I have neglected to argue for "the substantial ontological assumptions" that 0 and the infinitely many other numbers are abstracta. But on this construal, the objection is still off the mark. I cannot have neglected to argue for such assumptions because, as Benacerraf already grants that numbers exist, that we have knowledge of them, and that that knowledge is of abstracta, there is nothing to argue for. Benacerraf assumes a realist ontology for the sake of an argument against the possibility of a rationalist epistemology for mathematics. My reply is an explanation of how mathematical knowledge is possible supposing its objects to be abstracta. Hence, Azzouni's criticism that there is no argument for "the substantial ontological assumptions arising in mathematics" misses the point. No argument is required because my reply is a reply to an *epistemic* rather than an ontological challenge to mathematical realism.

Azzouni says that I have not provided "the needed dependability with respect to the ontologically independent abstracta," and, hence,

> ...those who want *knowledge* still have the worry, with respect to existentially-committing mathematical principles, of *why* the epistemic processes described guarantee either the existence of [abstracta] or the properties those principles attribute to them (*ibid.*, p. 237).

He expands the point:

> This worry is open to an ambiguity that may allow someone to dismiss it as merely raising a sceptical concern.... To say merely that the best methods used by mathematicians are knowledge-yielding (because after all, they are the best *mathematical* methods) is to use an argument crystal-ball readers can also avail themselves of (*ibid.*, p. 237, footnote 24).

But I do not dismiss Azzouni's worry as a skeptical doubt. Rather, my criticism is that he does not take into account my reply to Benacerraf's challenge: he ignores my explanation of how mathematicians can have knowledge of mathematical objects to which they can have no causal access. My explanation is that "the needed dependability"—in Benacerrafian lingo "the restriction of the unwanted variation"—derives from the ability of reasoning in mathematical proofs to exclude all suppositions except the one the theorem expresses, in virtue of establishing truths about the intrinsic properties of objects that have their intrinsic properties necessarily. Azzouni's failure to address my account of the warrant for mathematical knowledge might suggest that his worry is in fact a skeptical worry about whether mathematical knowledge can have a warrant. This suggestion would both explain why Azzouni ignores my account and why he claims that I have "merely unearthed the challenge Kant used to bury [Cartesian and Leibnizian] rationalism in the first place" (*ibid.*, p. 237). Nonetheless, I do not think skepticism is what is behind Azzouni's worry.

To see what is behind Azzouni's worry, we need to focus on his comparison of my account of reasoning in mathematics with crystal-ball gazing. The undependability of crystal-ball gazing derives from the absence of a causal connection between what crystal-ball gazers claim on the basis of looking into their crystal balls and what exists in physical reality. Given this, Azzouni's criticism is that the undependability of our epistemology for mathematics derives from the absence of a causal connection between what mathematicians claim on the basis of constructing a proof and what exists in mathematical reality. The crystal-ball analogy, plus Azzouni's reiteration that the choice facing the realist is a contact epistemology or no epistemology at all, shows that he is in the grip of the causal theory. Nothing that the realist could say, short of caving in on the issue of "[causal] responsive[ness] to the mind-independent mathematical objects," would make epistemic sense to him. Thus, the possibility of a no-contact epistemology is never seriously considered.

*I.B.2. Mathematical knowledge: inferential steps.* Next, I turn to objections to what my rationalist epistemology says about inferential steps. Gary Malinas[16] objects:

> Katz's proposals for a rationalist epistemology in which the concept of proof occupies a central role faces a "justification problem." Which standards of proof are salient in evaluating *a priori* knowledge claims? Either there is One True Logic, or...*a priori* knowledge claims need to be relativized to background logics (*ibid.*, p. 427).

Malinas's question is irrelevant to the epistemology that constitutes my reply to Benacerraf's epistemic challenge, but not because there is no rationalist/empiricist controversy about logical knowledge nor because my rationalist epistemology does not apply to logic. There is such a controversy, and, in virtue of the fact that logical realism takes truths of logic to be necessary truths, my epistemology for mathematical knowledge can be adapted to logical knowledge. The irrelevance is due to something else.

Malinas's question is equivocal between the (philosophical) question of how the logician's best theory can be defended against philosophical skepticism and the (scientific) question of what theory logicians should adopt as the best theory of logical implication. But neither of these questions is relevant to my rationalist epistemology. The skeptical question is irrelevant for the same reason that philosophical skepticism is irrelevant to my reply to Benacerraf. Rationalism and empiricism are second-level theories, so the issue in the controversy between them is which provides the best understanding of knowledge in logic. As indicated, it is a condition of the controversy that we have the first-level knowledge in question.[17]

The scientific question is irrelevant because a theory of knowledge is not part of logic proper. An epistemology for logic is a piece of the philosophy of logic, not logic, just as an epistemology for mathematics is a piece of the philosophy of mathematics, not mathematics. Neither Benacerraf's empiricist epistemology nor my rationalist epistemology addresses questions in mathematics proper, such as whether there are alternative pure geometries or "One True Geometry."[18] By the same token, neither addresses questions in logic proper. The choice among alternative logics or the question whether there is "One True Logic" falls within logic and is to be decided by logicians, not by philosophers. Hence, Malinas's question is irrelevant to a rationalist epistemology: the first interpretation belongs to the foundations of the sciences and the second falls within a science at the first level.

---

[16] "Review of *Realistic Rationalism*," *Studia Logica*, LXIV, 3 (2000): 425-29.

[17] Benacerraf's ("Mathematical Truth," p. 413) own empiricist epistemology, and indeed any serious epistemology for scientific knowledge, is vulnerable to philosophical skepticism because its account of scientific knowledge of the past, laws, and theories must assume a fixed system of inference rules to get beyond observation statements about medium-sized objects in the present. Accordingly, my explanation of why there does not have to be causal contact with mathematical objects faces no "justification problem" of a skeptical nature in connection with its assumption of a fixed system of logic.

To be sure, there is a "justification problem" here, but it is the quite different problem of showing that my rationalist epistemology provides a better philosophical understanding than empiricism of logical and mathematical knowledge. The problem became pressing to modern-day rationalists with the appearance of Quine's[19] holistic empiricism. Until then, rationalists in logic and mathematics had little cause for concern. Gottlob Frege had dealt with J. S. Mill's inductivism and Quine with Carnap's conventionalism.[20] But, having neither the implausibility of Mill's inductivism nor the emptiness of Carnap's conventionalism, Quine's holistic empiricism gave empiricism a new lease on life. The challenge posed by this new empiricism is our justification problem.

The problem is solved if Quine's holistic empiricism can be refuted. Earlier, we saw a counterexample to Quine's holism reflecting the tension between its mathematical realism and its uncompromising empiricism. There is another tension, one between its constitutive principles and its uncompromising empiricism.[21] If *everything* in our conceptual scheme is revisable in principle, then the constitutive principles of Quine's holistic empiricism are revisable in principle. In particular, the principle of principle revision which tells us that we

---

[18] It is well to recall just how difficult the discovery of non-Euclidean geometry actually was. From the time of Proclus to the seventeenth century, mathematicians tried to derive the parallel postulate from the other Euclidean axioms. In the sixteenth century, Giovanni Girolamo Saccheri and Johann Heinrich Lambert tried to prove the postulate by reductio, deriving "absurdities" which, unrecognized by them, were theorems of non-Euclidean geometry It took Johann Carl Friedrich Gauss, János Bolyai, and Nikolai Ivanovich Lobachevsky to show that we can have a perfectly consistent geometry based on a non-Euclidean system of axioms. For further discussion, see Richard Courant and Herbert Robbins, *What Is Mathematics?* (New York: Oxford, 1941), pp. 218-22.

[19] *Word and Object* (Cambridge: MIT, 1960), especially pp. 275-76.

[20] Frege, *Translations from the Philosophical Writings of Gottlob Frege*, Peter Geach and Max Black, eds. (New York: Philosophical Library, 1952), see also *The Foundations of Arithmetic*, J.L. Austin, trans. (New York: Philosophical Library, 1953); Quine, "Truth by Convention," in *The Ways of Paradox and Other Essays* (Cambridge: Harvard, 1976), pp. 77-106.

[21] See my *Realistic Rationalism*, pp. 67-74.

"must"[22] reevaluate to restore consistency with experience is revisable in principle. But, when we stop to think about it, the whole idea of revising a principle in accord with itself smacks of paradox. How can we discredit the principle when doing so would discredit the argument for discrediting it? A reevaluation of the principle requires a sound argument showing that changing its truth value to falsehood is the best way of removing a conflict with experience. Since the principle itself figures as a premise in the argument, the argument cannot be sound. If the premise is true, assuming the argument is valid, it is not an argument for reevaluating the principle, since then the conclusion is true, too. If the premise is false, the argument is unsound. Given that the principle is not revisable and the principle that everything is revisable, Quine's holistic empiricism is inconsistent.[23]

*I.B.3. Mathematical knowledge: possession premises.* How can statements be known? We can distinguish three classes of possession premises. First, there are the statements that can be proved as theorems. For them, what I have already said suffices. Second, there are statements—perhaps Church's thesis is an example—that are too highly theoretical for proof. I have argued elsewhere[24] that such statements can be known a priori on the basis of a priori demonstrations of their indispensability to the best systematization of their branch of mathematics. This can suffice for present purposes. Third, there are statements, like '4 is composite', that are too basic for proof and too highly untheoretical for indispensability arguments. I shall devote the remainder of this section to explaining how this third type of statement is known.

---

[22] Quine's term—see *From a Logical Point of View*, p. 42

[23] Quine's holistic empiricism rests on his argument that concepts in the theory of meaning, such as analyticity, on which philosophers have relied to explain the a priori in logic and mathematics, cannot be made objective sense of (see footnote 44 below). I argued that this argument overlooks the most natural way to make objective sense of them, namely, providing an axiomatic theory that defines them in terms of their relations to one another, which is the way concepts are defined in Peano arithmetic, predicate logic, and generative grammar—*The Metaphysics of Meaning* (Cambridge: MIT, 1990), pp. 175-93. What is more, Quine himself now concedes this point—see Chalmers Clark, "Quine and the Speculative Reach of Philosophy in Natural Science," *Dialectica*, LII (1998): 275-90, esp. pp. 280-81.

Moreover, to suppose that the philosopher's task is fundamentally different from that of the scientist does not, contrary to Quine, imply "a vantage point outside the conceptual scheme he takes in charge"—*Word and Object*, p. 275. It supposes only that the differences between natural science, formal science, and philosophy, which form parts of the scheme, are differences in kind, not degree. As we saw in connection with (Q1) and (Q2), the differences in kind derive from the nature of the subjects.

To be sure, logician-philosophers sometimes inject philosophical considerations into arguments in logic, but one ought to take a dim view of this. Resting decisions about a matter of logic on considerations outside of logic proper is a dubious practice not only because it introduces external considerations, and typically highly contentious ones, but also because it leaves us with no nonpartisan theory of implication to appeal to in determining which of the competing philosophies of logic offers the best account of our logical knowledge.

[24] *Realistic Rationalism*, pp. 47-48.

Realists say that this third type of statement is known by intuition. I think they are right. As indicated above, the claims of antirealists that appealing to intuition is "mysticism" have force only if Benacerraf's dilemma is firmly in place. For once we have the alternative of a no-contact epistemology available, antirealists can no longer assume that the realist has to choose between a natural connection and incoherence or an unnatural connection and incredibility. The immediate task, therefore, is to explain how intuition fits into my no-contact epistemology.

To be sure, there are philosophers who label intuition mystical without seeming to base the charge on Benacerraf's dilemma. For example, Resnik[25] says he "finds intuition mysterious" (*ibid.*, p. 210). But, once Benacerraf's epistemic challenge is met, it is no longer clear what the mystery is or why the mystery, whatever they think it is, is supposed to matter philosophically. Resnik, for example, does not say why intuition is more of a mystery than thinking or imagining. It is generally recognized that psychology cannot tell us much about the workings of thought and imagination, yet we do not hear the cry "mystical" or "mysterious" raised about them. Why the double standard?

One further point in this connection. Nothing stops realists from taking the view that intuition is a natural faculty like thinking and imagining. Their commitment to the abstract being something over and above the concrete entails no commitment to the mental being something over and above the physical. Not only is there no commitment to mind/body dualism, but abstract/concrete dualism raises none of the problems of the former because there is no putative causal cross-over to explain, that is, no analogue to robust psychophysical correlations. Further, taking a naturalistic view of the process of intuition does not prevent one from claiming that intuition provides an a priori warrant for basic mathematical truths. "Mystical" and "mysterious" are inflated rhetoric.

The other factor that has encouraged philosophers to think that intuition cannot play the epistemic role in which realists traditionally cast it is the plethora of mischaracterizations of intuition. Perhaps the most influential mischaracterization is Wittgenstein's description in the *Philosophical Investigations*[26] of intuition as "an inner voice" (*ibid.*, section 213). Given the context (*ibid.*, section 214), it is clear that this description is intended to get the reader to take intuitions as linguistic phenomena so that they become subject to the "variety of interpretations" that give rise to the rule-following paradox. But having a mathematical intuition is not at all like hearing an inner voice. The paradigm of an inner voice is when I say to myself, 'Don't forget to pick up a pizza tonight or you're in big trouble'. The intuition we have of the truth of a mathematical principle—for example, the pigeon-hole principle—is nothing like this. There is nothing linguistic about such an intuition; it is a purely intellectual grasp of mathematical structure. Wittgenstein's characterization is a phenomenological category mistake.

---

[25] "Review of *Realistic Rationalism*," this JOURNAL, XCVI, 4 (April 1999), 207-11.

Wittgenstein also called intuitions "an unnecessary shuffle": "if it can guide me right, it can also guide me wrong" (*ibid.*, section 213). (See also the more recent form of this criticism due to Philip Kitcher.[27]) But if there is no threat of the rule-following paradox, this criticism comes down to no more than the jejune claim that intuitions are fallible. This is not something that should worry us. We recognize that fallibility is compatible with providing a warrant for knowledge in the case of observation and inference, so that, without some relevant difference between them and intuition, we ought to recognize this compatibility in the case of intuition, too. In fact, there is no difference between them with respect to checking results. Intuitions about which we have doubts can be checked against a systematic theory developed on the basis of a wide range of intuitions about which we have no doubts. Hence, as with observation and inference, we can discover errant intuitions and revise them on the basis of what such a theory says about particular cases.

Since theoretical systematization in mathematics is an a priori process, revision of errant intuitions is revision on a priori grounds. Of course, some antirealists deny that revision of intuitions is only for a priori reasons, citing alleged examples of the revision of intuitions based on empirical evidence. The examples include the "intuition" that the earth is flat, that its geometry is Euclidean, and that this or that person has such-and-such an inner state. On the one hand, it is not clear why such cases ought to be accepted as examples of intuition in the appropriate sense. The judgment that the earth is flat is just a bad inference from immediate experience. The judgment that Euclidean geometry is true in the world is not a judgment about pure mathematics, since it makes no sense to say that pure Euclidean geometry, where the notions of point and line are uninterpreted, is physically true. An applied Euclidean geometry may, of course, be true or false, but it is not part of pure mathematics. Finally, the term 'intuition' that applies to judgments that persons have such-and-such feelings or motives is no more than a homonym.

---

[26] G. E. M. Anscombe, trans. (New York: Macmillan, 1953).

[27] *The Nature of Mathematical Knowledge* (New York: Oxford, 1984), p. 64.

On the other hand, no example can be accepted even as prima facie relevant to the issue without a reason to think it falls within the scope of intuition in the appropriate sense, and the only way we can determine its scope is on the basis of an explication of the notion "intuition." Thus, relevance presupposes that intuition itself has been made a subject of study and explicated in the same way as notions like "argument" in logic and "sentence" in linguistics. Only at the point where significant progress toward an explication of intuition has been made, a point we are as yet nowhere near, can we say with any confidence what is a case of intuition. As things now stand, the alleged counterexamples are quite premature.

Further, since explication starts with clear cases central to the explicandum, since the central clear cases for the explication of intuition are mathematical and logical cases, and since it is plausible to think they are a priori, it is plausible to think that the extension of the explicans will contain exclusively a priori intuitions. Hence, nothing a posteriori will turn out to count as cases of intuition, and none of the examples cited as counterexamples will be relevant to the issue of a priori revisability.

Mischaracterizations of intuition are found among its friends as well as among its enemies. The misrepresentations on the part of its friends are often the most harmful because they are naturally assumed to be presenting intuition in the best possible light. Noam Chomsky[28] is a case in point. He, more than anyone, is responsible for the revival of interest in intuition in linguistics and the philosophy of language over the last fifty years, but his psychologism leads him to mischaracterize linguistic intuitions as informing us about the structures in the mind/brain which comprise the speaker's knowledge of the language rather than about the language itself (*ibid.*, pp. 18-27).[29]

---

[28] *Aspects of the Theory of Syntax* (Cambridge: MIT, 1965).

[29] Explicit knowledge as well implicit knowledge, since, on Chomsky's view, truth in the former case lies in correspondence to the implicit knowledge. Chomsky writes: "...the grammar [a theory of a language] is justified to the extent that it correctly describes its object, namely the linguistic intuition—the tacit compe-

Accordingly, the intuition that the sentence 'Visiting relatives can be annoying' is ambiguous does not present us with a fact about the sentence, but with a fact about speakers. Hence, intuitions provide no a priori warrant for claims about sentences and languages (collections of sentences), but only an a posteriori warrant for claims about speakers' (tacit) beliefs about sentences and languages.

Chomsky thinks that aspects of the linguistic competence of speakers have epistemic force because a significant number of the principles comprising this competence are innate and hence can be accorded species-universality. Such thinking underlies much of the interest in the issue of nativism in philosophy, but while nativism is a significant issue in psychology, particularly in the study of learning, it has no direct epistemological significance. Origins, even species-universal ones, carry no normative force. Discovering that we are innately programmed to think that there are only finitely many English sentences would not establish that there are only finitely many English sentences. Similarly, discovering that we are programmed to think that there are more numbers than even numbers or that space is Euclidean would not establish these propositions. Chomsky confuses nativism, which is a scientific doctrine about the origin of the speaker's tacit knowledge of the language, with rationalism, which inter alia is a philosophical doctrine about the epistemic status of the linguist's explicit knowledge of the language.

Bealer, unlike Chomsky, takes intuition to be a source of a priori knowledge. He nevertheless also characterizes it in a way that sacrifices its ability to provide an a priori warrant. Bealer says that "to have an intuition that A is just for it to *seem* to you that A" (*op. cit.*, p. 123). But if intuitions were seemings, they would present us with nothing more than an appearance of mathematical reality and we would face the problem of discovering what the mathematical reality is behind the appearances. The problem cannot be solved by taking seemings to have a causal relation to their objects, in the manner of sensory perceptions, because mathematical objects are abstract and sensory perception is a causal process. If seemings have no causal relation to the mathematical reality of which they are appearances, what ensures that a seeming with which we have an inner acquaintance represents the mathematical reality outside us as it is?

Bealer's answer is that appearances warrant knowledge of mathematical objects because they can be counted as evidence in virtue of

---

tence—of the native speaker. In this sense, [it] is justified...on grounds of correspondence to linguistic fact" (p. 27).

a "tie to the truth" (*op. cit.*, pp. 122-30). This explanation is inadequate for two reasons. First, generally speaking, evidence does not provide mathematical justification. An intuition that gives us the solution to a mathematical problem (for example, a tiling problem) is not evidence for the solution. It proves it. In natural science, there is increase in confirmation with increase in evidence, but not in mathematics. Increasing the number of cases where Goldbach's conjecture is confirmed, however large the increase, takes us not one scintilla closer to establishing that the conjecture is true.

Moreover, that there is such a difference between the natural and formal sciences is only to be expected on this distinction between them. If the natural sciences use observation of concrete objects to prune down what is possible about them to what is actual about them, then ceteris paribus increasing confirmation will increase the warrant we have. If the formal sciences use reasoning about abstract objects to prune down what is supposable about them to what is necessary about them, then simply increasing confirmation—adding more cases where Goldbach's conjecture is confirmed—will do nothing to provide a warrant. A warrant requires ruling out all contrary suppositions, but, unlike the case of concrete objects with causal relations among them, there is nothing here that enables us to say something about unexcluded suppositions on the basis of excluded ones.[30]

Second, Bealer's explanation of "a tie to the truth" fails to provide a rationale for taking the way mathematical objects seem to us to be as a warrant for claims about the way they actually are. He tells us that the tie is a "modal tie" (*op. cit.*, p. 130). He says that, if we go through a process of "a priori justification"—canvassing intuitions, critiquing them, constructing theories that systematize the surviving intuitions, testing the theories with more intuition, and repeating the process until reflective equilibrium—then we can be assured that most of the propositions at the equilibrium stage of the process are true. But this account of "a tie to the truth" is question-begging. The process of a priori justification is supposed to explain the evidential status of intuition, but it turns out to depend on intuitions that are already assumed to have evidential status (that is, to have a tie to the truth).

---

[30] Surprisingly, Gödel says: "If mathematics describes an objective world.. there is no reason why inductive methods should not be applied in mathematics"—"Some Basic Theorems on the Foundations of Mathematics and Their Implications," p. 313 Gödel does not tell us what form they would take, though he gives an example of estimating probabilities, one which to my way of thinking—and to George Boolos's (*Logic, Logic, and Logic* (Cambridge: Harvard, 1998), p. 112)—does not help clarify things In any case, if my distinction between the formal and natural sciences is right, there is a reason.

Hence, Bealer's account fails to explain how a seeming about numbers relates to the objective truth about them. Without more than a tie to truths about the contents of our minds, we are in the dark as to why intuitions warrant claims about an independent mathematical reality.

The principal obstacle to fitting intuition into my no-contact epistemology is the widely accepted view that intuition is separate from reason. It is true that intuition and reason commonly play different roles in mathematical proof, the former warranting the truth of premises and the latter warranting the validity of the steps from premises to conclusion, but this difference does not require taking them to be separate faculties. Intuition and reason can be taken to be different functions of one and the same faculty. We can say that reason is rationality in application to deductive structures and intuition is the same faculty in application to elements of such structures. We can think of intuition as reason in the structurally degenerate case.

Here is an argument that intuition is part of reason. Each of the finitely many steps in a proof is either minimal or analyzable into substeps. If the step is minimal, the inference is immediate, in which case it corresponds to an element in the deductive structure, taking the form of a proposition expressing a direct implication. In this case, the validity of the step comes down to the truth of this proposition and ex hypothesi its truth must be sanctioned by the rational faculty under the aspect of intuition. If the step is analyzable into further steps, the same question will arise about each of them: Is the step minimal or is it analyzable? Since the reasoning in question is a proof, there is no infinite regress, and eventually we reach minimal steps. Since the inference in these cases is also immediate, the sanction of reason in each appears again under the aspect of an intuition. The sequence of steps presents itself as a concatenation of the intuitions.

This double-aspect theory of reason has two important consequences. One is that the theory provides a far stronger argument for intuition than has thus far been available. Until now, the argument for intuition had to be made on the basis of examples where intuition is the best explanation of our immediate knowledge of the truth of a principle.[31] Given that intuition is part of reason, philosophers either have to reject both intuition and reason as mystical, or accept both as aspects of one and the same natural faculty. Thus, we have the more compelling argument: dismissing intuition is dismissing reason.

---

[31] See my *Realistic Rationalism*, pp. 43-47.

The other consequence of the double-aspect theory is that it enables us to provide a better characterization of mathematical intuition. Carrying the earlier characterization of reason over to intuition, we can say that mathematical intuition is an immediate apprehension of a mathematical structure that excludes all possibility of mathematical objects with that structure not having a certain mathematical property. Hence, knowledge that 4 is composite is based on an immediate apprehension that its arithmetic structure excludes all possibility of its not being composite.[32] This apprehension is of 4's *being* composite, not of its seeming so and the seeming, in virtue of some further epistemic process, serving as evidence for (some probability of) its being composite.

Mathematical intuitions involve some sort of mental representation. The representation functions to focus our thoughts on the mathematical object(s) in question, that is, the number 4 in the above example. Hence, intuition presents us with knowledge of something objective, in contrast to introspection, which presents us with something subjective. As Frege would put it, one person's introspective object is not that of another (*op. cit.*, p. 59). Since apprehending a seeming causally relates us to something subjective, what Bealer calls "intuition" is really introspection. Further, intuition differs from introspection with respect to the fact that the relation of the former to its object is noncausal while the relation of the latter to its object is causal. The relation between the representation and the object in intuition is noncausal in the manner of the correspondence between mathematical truths and their abstract truth makers.

This characterization of intuition enables me to rebut Kitcher's criticism of intuition. He claims that neither of the two ways intuition can be understood provides a warrant for mathematical beliefs (*op. cit.*, p. 61). One is as a heuristic insight that offers "a fruitful gestalt on a problem." Such intuitions are, he says, a real part of mathematical practice, but at best only "a *prelude* to mathematical knowledge." I agree. In addition to distinguishing mathematical intuition from introspection, I also want to distinguish it from mathematical insight, and we can do so on the grounds that mathematical intuition is a mode of justification while mathematical insight is a mode of discovery. The other way Kitcher thinks intuition might be understood is as "gazing on the mathematical objects and coming up with a fruitful idea." "Such intuitions," he says, "are phantasies" (*op. cit.*, p. 61). Here, Kitcher is assuming that realists are forced to adopt a perception-based epistemology. Thus, like the other antirealists we have considered, he overlooks the possibility of a no-contact epistemology. In such an epistemology, intuition, as I have characterized it above, is a mode of justification rather than a mode of discovery, a real part of mathematical practice, a natural faculty, and a warrant for mathematical knowledge in cases where propositions are too basic for proof and too untheoretical for indispensibility arguments.

---

[32] My characterization of mathematical intuition applies straightforwardly to intuition in the other formal sciences. On my realism, the abstract objects studied in logic and linguistics differ from the objects studied in mathematics only in having logical and linguistic structures rather than mathematical structures.

## II. EXAMPLES OF SEPARATISM

We saw that the advantages of integrating realism and rationalism come from the support each receives from the other: rationalism enables realism to meet Benacerraf's epistemic challenge and realism enables rationalism to explain how reason can provide a full account of mathematical knowledge. In the present section, I want to argue that, correspondingly, the disadvantages of "realism without rationalism" and "rationalism without realism" come from each needing support from the other which their separatism denies them.

Gödel's position is realism without rationalism. It is mathematical realism based on conceptual realism.[33] Gödel says: "concepts form an objective reality of their own, which we cannot create or change, but only perceive and describe."[34] His position might also be characterized as mathematical realism with Husserlian phenomenology, as he based his notion of perceiving mathematical concepts on Edmund Husserl's conception of the intuition of essences. As Richard Tieszen[35] observes, Gödel stresses the similarities between the intuition of mathematical concepts and the perception of physical objects. And

---

[33] Gödel's conceptual realism: "a mathematical proposition...is true already owing to the meaning of the terms occurring in it, irrespectively of the world of real things"—"Some Basic Theorems on the Foundations of Mathematics and Their Implications," p. 320. Boolos points out that Gödel neglects to address the possibility that "the axioms of set theory [are] true, not in virtue of the concept of a set..., but simply because sets just happen to be as the axioms have it" (pp. 117-18). "Why," he goes on to ask, "must our knowledge of sets be mediated solely through our understanding of the concept of a set...?" An answer might be that Gödel thought Husserlian intuition, which he hoped would provide him an epistemology, was appropriate for concepts. But, since, for Gödel, concepts are abstracta every bit as much as sets, the epistemic problem for them is the same. Hence, as explained in the text, it is unsolvable along Husserlian lines. This, rather than the Quinean issues that Boolos raises, is where Gödel's conceptual realism runs into trouble.

[34] "Some Basic Theorems on the Foundations of Mathematics and Their Implications," p. 320.

[35] "Kurt Gödel and Phenomenology," *Philosophy of Science*, LIX, 2 (1992): 176-94, see pp. 180-81.

Hao Wang[36] reports Gödel saying: "The analog of perceiving sense objects from different angles is the perception of different logically equivalent concepts" (*ibid.*, p. 233). That is, clarification of the content of a mathematical concept involves seeing it from the perspective of different logically equivalent concepts in a way analogous to the way Husserl takes clarification of a physical concept to involve seeing its instances from different visual perspectives. The mathematical concept is subjected to "free variation" in Husserl's sense: we vary aspects of the concept to see which properties are essential. The essential properties are those which remain invariant, the ones which belong to all the logically equivalent concepts which arise under such variation.

There are two problems with Gödel's position. One is the familiar problem that Benacerraf raises about Gödel's general epistemology. Gödel[37] says mathematicians perceive abstract objects, but he makes it clear that such perceptions "cannot be associated with actions of certain things upon our sense organs" (*ibid.*, pp. 267-68). Gödel says that what he has in mind is "another kind of relation between ourselves and reality" (*ibid.*, p. 269), but he never says what that is.[38] Benacerraf[39] properly objects that, since Gödel speaks only figuratively, the analogy with actual perception is too "superficial" to provide epistemological insight into mathematics. Now, as concepts are abstracta for Gödel,[40] his account of the intuition of concepts is subject to Benacerraf's objection. Gödel has no account of the intuition of abstract concepts.

The other problem is that a determination of the essence of a concept on the basis of logically equivalent concepts is too coarse-grained a criterion for concept identity. The concept 'the number 2' and the concept 'the number which is the only even prime' are logically equivalent, but they are not, as it were, aspects of one concept perceived from different angles. Rather than essential aspects of one and the same mathematical concept, '2' and 'the only even prime' are two independent but equivalent concepts, as shown by the fact that we have to prove that 2 is the only even prime.

---

[36] *A Logical Inquiry: From Gödel to Philosophy* (Cambridge: MIT, 1996).

[37] "What Is Cantor's Continuum Problem?"

[38] What to make of Gödel's statement that "I believe there is a causal connection in the perception of concepts" (Wang, p. 235)? The tiny bit of exposition ("Already *noûs* in Aristotle is a causal affair") does not make matters easier to understand. It is thus preferable to treat such a rare remark as a lapse.

[39] "Mathematical Truth," pp. 412-16.

[40] See also Wang, p. 235, section 7.3.10.

Shapiro's position, similarly, is realism without rationalism. His epistemology sounds like Gödel's:

> With processes much like—or even identical to—ordinary sense perception, a subject comes to recognize and learn about patterns.... We do not literally see or hear abstract entities. Our subject comes to recognize a pattern by observing patterned systems (*op. cit.*, p. 111).

According to Shapiro, our knowledge of "the set theoretic hierarchy" is based on perception of concrete instances of small (finite) structures, but also requires the exercise of linguistic abstraction and definition (*op. cit.*, p. 112). Even without raising questions about whether linguistic abstraction and definition are up to the task Shapiro assigns them, it is clear that his epistemology is subject to Benacerraf's criticism of Gödel's. For, Shapiro's notion of "recogniz-[ing] a pattern" is either intended figuratively—the emphasis is on "much like"—or intended literally—the emphasis is on "identical to." In the former case, the analogy is too "superficial" to provide an epistemology for mathematical knowledge, while, in the latter, what is said about mathematical knowledge, that it is based on perception of the abstract in the concrete, is impossible.

Resnick's realism is realism with empiricism. His epistemology is a Quinean holistic empiricism on which "[w]e can empirically test certain hypotheses about numbers—modulo an appropriate set of auxiliary hypotheses—but the bulk of our evidence for numbers has been the success of the framework positing them."[41] Since, as we saw above, Quine's holistic empiricism is untenable, Resnik's epistemology is untenable as well.

Bealer's position is rationalism without realism. The absence of realism in his position leads, as argued already, to an account of intuition that attempts to put the square peg of an epistemology for knowledge of the concrete into the round hole where an epistemology for knowledge of the abstract should go.

Nagel's position in *The Last Word* (*op. cit.*) is also a case of rationalism without realism. In his rationalist defense of mathematical knowledge against skepticism and empiricism, Nagel argues that we cannot

> ...rise *above* the logical thought [for example, that 2 + 2 = 4], thinking about it while withholding commitment from its content.... [The thought that 2 + 3 may be 4 or 3 or something else] is unintelligible, for two reasons. First, it includes the "thought" that perhaps 2 + 3 = 4, which has not been given a sense and cannot acquire one by being

---

[41] Resnik, *Mathematics as a Science of Patterns*, p. 111.

> conjoined with the extraneous, non-arithmetical thought that an evil demon might be manipulating [one's] mind. Second, [skeptical reasoning] is itself an exercise of reason, and by engaging in it Descartes has already implicitly displayed his unshakeable attachment to first-order logical thought... (*op. cit.*, pp. 59-60).

Neither reason fills the bill.

First, as Nagel notes, arithmetic falsehoods like '2 + 3 = 4' "are not gibberish. [2 + 3 = 4] has enough sense to be necessarily false, and it can enter into reasoning as the premise or conclusion of a reductio ad absurdum" (*op. cit.*, p. 59, footnote 6). To counter this problem, Nagel says: "Nevertheless, though one can *suppose* for the purpose of argument that 2 + 3 = 4, or observe that it follows from certain assumptions that 2 + 3 = 4, it is not possible to *think* that (perhaps) 2 + 3 = 4" (*op. cit.*, p. 59, footnote 6). But that begs the question: Nagel is trying to show what he now states without argument, that "it is not possible to *think* that (perhaps) 2 + 3 = 4."

Second, a skeptic like Descartes or an empiricist like Quine may acknowledge "an unshakeable attachment to first-order logical thought" without thereby "implicitly ruling out [the possibility that 2 + 3 = 4]." An attachment to "first-order logical thought" does not imply a (self-refuting) attachment to arithmetic truth. Frege's attempt to reduce mathematics to logic, more than first-order logical thought in any case, failed.

Without raising the question of how Nagel might reformulate his claim, I think it is clear that both Descartes and Quine offer us ways of explaining how we can think an allegedly necessary truth, such as an arithmetic truth, is false. Descartes's suggestion is based on his notion of the absence of any limit on God's power, Quine's on his conception of how logical and mathematical knowledge might be relativized to observation. Descartes's way is much trickier, so I shall describe Quine's.[42]

---

[42] For a discussion of Descartes, see my *Realistic Rationalism*, pp. 53-55. A further point needs to be made here. Nagel says against Descartes: "I cannot think... that I would be in an epistemically identical situation if 2 + 2 equaled 5 but my brains were scrambled—because I cannot conceive of 2 + 2 being equal to 5" (*op. cit.*, p. 63). There is a subtle mistake here. Besides the sense of directly conceiving, there is also the broader sense of conceiving. On neither can I conceive gibberish such as that two plus two is Hesperus, but on the latter, unlike the former, I can conceive that 2 plus 2 is 5. What I have to do is to find a way of explaining how it might be so in spite of my inability to directly conceive its being so. But conception in this sense is all that is required for Descartes's skepticism, since it requires no more than an account of how it might turn out that a proposition that we cannot directly conceive to be true might nonetheless be true. This is what Descartes takes himself to have accomplished in reminding us of God's omnipotence. Thus, Nagel's

Of course, I have to revise Quine's conception to avoid my earlier criticism. To do this, I drop the mathematical realism and exempt the constitutive principles of his holistic empiricism from revisability. On such a revised empiricism, we come to know that an arithmetic statement is false in the same way that we come to know that an empirical statement is false: by showing that assigning it falsehood is the simplest redistribution of truth values that removes certain conflicts with experience. Hence, there is a basis for claiming that we can think '2 + 2 = 4' might express a contingently false statement, as well as a basis for explaining why we find it hard to conceive of the falsehood of logical and mathematical truths, namely, their centrality would require impugning too much of the whole system.

## III. METAPHILOSOPHY

Wittgenstein's and Quine's metaphilosophies are based on accounts of knowledge as exclusively natural knowledge. Wittgenstein's[43] ascends to his linguistic positivism from his naturalist account of knowledge which, as originally set out in the *Tractatus*, restricts meaningful language to the "propositions of natural science" (*ibid.*, section 6.53). The naturalism of the *Tractatus* lives on in the *Philosophical Investigations*, where Wittgenstein says that philosophy should be no more than "remarks on the natural history of human beings" (*op. cit.*, section 415). Such remarks are in the service of his new therapeutic philosophy which is designed to rid us of philosophical problems by "bring[ing] words back from their metaphysical to their everyday use" (*op. cit.*, section 116). The positivism of the *Tractatus*, too, lives on in the *Philosophical Investigations*, where Wittgenstein writes: "The results of philosophy are the uncovering of one or another piece of plain nonsense and of bumps that the understanding has got by running its head up against *the limits of language*" (*op. cit.*, section 119, italics added).

Quine ascends to his holistic empiricism from an account of knowledge on which it is all natural knowledge. On his methodological naturalism, common sense and its "continuation," science, begin with sense experiences and aim at "expedit[ing] our dealings with sense experience." Observation of the natural world enters, to some degree, into the justification of every statement in "[t]he totality of our so-called knowledge or beliefs," including mathematics, logic, and

---

defense of rationalism does not meet the empiricist challenge to the absolute truth of logical and mathematical principles because, as I explain below, (revised) Quinean empiricism explains how we can conceive of 2 + 2 being 5, in spite of our being unable to conceive it directly.

[43] *Tractatus Logico-Philosophicus*, David F. Pears and Brian F. McGuiness, trans (New York. Harcourt Brace, 1922).

philosophy. Philosophy, he says, is "an enterprise within natural science."[44]

The starting point for Wittgenstein's and Quine's ascents is an exclusively naturalist account of knowledge. In contrast, the starting point for ascent here is a non-naturalist account of one branch of knowledge, mathematics, which, as a formal science, differs from natural science both in the questions it asks and in the methods it uses to answer them. For Wittgenstein and Quine, mathematics is something natural—natural decision, in the former case, and natural science, in the latter.[45] On my account, mathematics is something non-natural, namely, the study of abstracta with mathematical structure. Contrary to Wittgenstein, mathematical knowledge is discovery, and, contrary to Quine, it is a priori.[46]

---

[44] Quine, "The Nature of Natural Knowledge," in Samuel Guttenplan, ed., *Mind and Language* (New York: Oxford, 1975), p. 68. The ascent to this metaphilosophy is set out in "Two Dogmas of Empiricism" (in *From a Logical Point of View*, pp. 20-46). It starts in section 5 with Quine's diagnosis of the mistake brought to light in sections 1-4 by his criticisms of the analytic/synthetic distinction (*From a Logical Point of View*, pp. 41-42). He takes it that the thought that "the truth of any individual statement" is somehow analyzable into "a linguistic component and a factual component" has been shown to be "nonsense, and the root of much nonsense." By a kind of reductio argument, he concludes that "science['s] double dependence upon language and experience...is not significantly traceable into the statements of science taken one by one. ...in taking the statement as unit we have drawn our grid too finely. The unit of empirical significance is the whole of science" (p. 42). The familiar holistic empiricism to which this conclusion leads is spelled out in section 6, pp. 42-46.

[45] Wittgenstein writes, "I am trying to say something like this: even if the proved mathematical proposition seems to point to a reality outside itself, still it only expresses acceptance of a new measure [of physical reality].... Why should I not say: in the proof we have won through to a *decision*?"—*Remarks on the Foundations of Mathematics* (Oxford: Blackwell, 1956), II, 27.

On Quine's holistic empiricism, mathematical objects are "posits" for the purpose of "expedit[ing] our dealings with sense experience": the "objective" of "total science, mathematical and natural and human," is the construction of simple laws for "predicting future experience in the light of past experience"—*From a Logical Point of View*, pp. 44-45.

[46] Note that I (*The Metaphysics of Meaning*, pp. 21-162; 175-202) have given reasons to think Wittgenstein's and Quine's arguments that all knowledge is natural knowledge are not adequate. The basic reason is that their arguments are too narrowly tailored to Fregean, Russellian, and Carnapian theories of meaning. Note also that more is necessary to block the route from the formal sciences to rationalist metaphilosophy than either a naturalism like Quine's which accepts realism or a naturalism like Wittgenstein's which rejects such large portions of mathematical practice. Quine's naturalism is the view I called ***methodological naturalism***, as distinct from both ***ontological naturalism***, which holds that all objects are natural objects, and ***epistemic naturalism***, which holds that all knowledge is of natural objects—*Realistic Rationalism*, p. xii. Methodological naturalism holds that all knowledge comes from investigations of natural objects. But what is the argument for this? Presumably, we have to restrict investigation to natural objects because abstract objects cannot be investigated directly, but, then, this reason must be supported by something like Ben-

My account of mathematical knowledge as a priori knowledge of abstracta is itself a piece of a priori philosophizing, as can be seen from the statement of the account in section I. If the account is correct, the results of this philosophizing are themselves a priori knowledge. Given that some scientific knowledge as well as some philosophical knowledge is a priori, the sweeping claim of philosophical naturalists like Wittgenstein and Quine that *all* knowledge is a posteriori natural knowledge is false.

To claim, contrary to Wittgenstein, that the results of philosophy are knowledge is not to say that philosophy has a domain of objects that, in the manner of the sciences, is its own special province. That philosophy differs from the sciences in this respect is the grain of truth in the naturalist ideology underlying the "linguistic turn." But, contrary to that ideology, not having its own domain does not mean that philosophy is linguistic and/or conceptual analysis with no higher ambition than serving as therapy for the philosophical community or as clarification for the scientific community. A discipline does not have to have its own objects to have its own problems. The traditional problems of philosophy do not arise from the attempt to understand some domain of objects, but from the attempt to understand common sense, scientific knowledge, and philosophical knowledge.

It seems reasonable to assume that knowledge of the foundations of the foundations of the sciences, if such philosophical knowledge can be had at all, is a priori philosophical knowledge. Given this, the main thesis of metaphilosophy is this:

> Knowledge in the formal sciences and in philosophy is a priori knowledge. The former is knowledge of abstract objects; the latter is not knowledge of objects at all, but either knowledge of how to best understand knowledge in the formal sciences or knowledge of the foundations of the foundations of scientific knowledge.

Quine's metaphilosophy transforms philosophy into science and Wittgenstein's does away with philosophy entirely. On my account, metaphilosophy leaves philosophy as it has always been in the work of the great philosophers: an autonomous discipline that seeks a priori

---

acerraf's epistemic criticism of realism. If so, our rationalist epistemology undercuts Quine's methodological naturalism.

Wang (p. 179) reports that Gödel lost "[h]is habitual calm" in connection with Wittgenstein's treatment of contradiction in *Remarks on the Foundations of Mathematics*: "Has Wittgenstein lost his mind? ...For example, 'You can't derive everything from a contradiction'. He ought to try to develop a system of logic in which that is true." Too bad there is no report of Gödel's reaction to Wittgenstein on Georg Cantor (see *Remarks*, pp. 54-58).

knowledge of the general aspects of being and knowledge. Such metaphilosophy provides philosophers with a job description other than "natural scientist" or "linguistic therapist." It is one on which philosophers are in Plato's and Aristotle's line of work, rather than, as the naturalizers would have it, in Darwin's or Freud's.

JERROLD J. KATZ

City University of New York/Graduate Center

## NOTES AND NEWS

The Editors regret the death of Jerrold J. Katz on February 7, 2002, at the age of 69. Professor Katz's contributions were broad, including linguistic semantics, philosophy of language, philosophy of linguistics, philosophy of mathematics, philosophy of science, metaphilosophy, and metaphysics. Born in 1932, he received his B.A. from George Washington University in 1954, served with the U.S. Army Counter Intelligence Corps from 1954 to 1956, and received his Ph.D. from Princeton in 1960. He first taught at the Massachusetts Institute of Technology before moving to the City University of New York Graduate Center in 1975, where he served as Distinguished Professor of Philosophy and Linguistics. Professor Katz, a pioneer in semantic theory and associated philosophical issues, wrote 10 books, co-wrote 1 book, edited 1 book, co-edited 2 books, and published more than 70 papers. His *Semantic Theory* (Harper and Row, 1972) had an enormous impact on philosophy, linguistics, psychology, and the fledgling field of cognitive science. Other works include: *Language and Other Abstract Objects* (1981), *Metaphysics of Meaning* (1990), and *Realistic Rationalism* (1998). On the day before his death, he signed a contract with Oxford University Press to publish his forthcoming book, *Sense, Reference, and Philosophy*. Among his awards and honors, he was a Senior Fellow at Harvard (1964-1965) and a Guggenheim Fellow (1972-1973), and he was presented with a Distinguished Alumni Scholar Award from George Washington University in 1989.

He was a man who, despite his distinguished career, did philosophy not for any career's prize, but to get things right by his own lights. He fashioned a highly original view of the nature of language and mathematics, and even when he wrote in a scholarly vein on figures such as Frege or Wittgenstein, it was always from the point of view that he himself had developed over the years. That he was often swimming against the stream of current tendencies in his subjects never shook him from his conviction in his ideas, or his determination to hone and refine them into the detailed and systematic philosophy he achieved by the end of his life.

# Philosophy

**NEW FROM CHICAGO**

## MAN MADE GOD

The Meaning of Life

**Luc Ferry**

*Translated by David Pellauer*

"Luc Ferry has taken a great risk in posing the only question that counts today, the question of the meaning of life. He stands out therefore as one of the most gifted philosophers of his generation. . . . Man Made God is a small step, but that of a giant."—Pascal Bruckner, *Le Nouvel Observateur*, on the French edition

PAPER $14.00

*Now in paper*

## IDENTITY AND DIFFERENCE

**Martin Heidegger**

*Translated by Joan Stambaugh*

"A welcome addition to the growing body of English translations of Heidegger. The special merit of Stambaugh's work is that the translator was a student of Heidegger's and was able to prepare this translation in consultation with him."—Stanley O. Hoerr and Staff, *The Review of Metaphysics*

PAPER $14.00

## ON TIME AND BEING

**Martin Heidegger**

*Translated by Joan Stambaugh*

"This lecture is perhaps the most significant document to be added to the Heideggerian corpus since the *Letter on Humanism*. . . . Stambaugh's translation is superb."—Stanley O. Hoerr and Staff, *The Review of Metaphysics*

PAPER $12.00

## BEING AFTER ROUSSEAU

Philosophy and Culture in Question

**Richard L. Velkley**

"I know of no more perceptive and learned commentator on the infrastructure of the development from Rousseau through Kant and German Idealism to Nietzsche and Heidegger than Richard Velkley . . . Essential reading for everyone who is interested in the ramifications of Rousseau's distinction between culture and civilization as well as in his grounding contribution to the modern conception of the historical nature of reason."—Stanley Rosen, Boston University

PAPER $18.00

## THE SLEEP OF REASON

Erotic Experience and Sexual Ethics in Ancient Greece and Rome

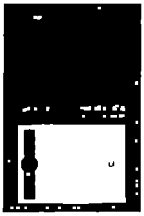
**Edited by Martha C. Nussbaum and Juha Sihvola**

"An outstanding collection of lucid essays whose contributors remain vigorously individual while resonating richly as an ensemble. The common thread in this work is a heightened attention to the ethical dimension of ancient thought on matters erotic, whose complexity the authors courageously and elegantly refuse to oversimplify."—Maud Gleason, author of *Making Men*

PAPER $26.00

THE UNIVERSITY OF CHICAGO PRESS

1427 East 60th Street, Chicago, IL 60637 • www.press.uchicago.edu

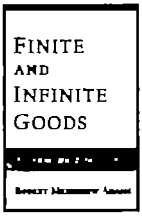


### VAGUENESS AND CONTRADICTION

**ROY SORENSEN**

Did Buddha become fat in one second? Is there a tallest short giraffe? According to Sorenson, the answers are yes! Although vagueness abounds in the way the world is divided, he argues that the divisions are sharp, yet we often do not know where they are. This book offers insight on language and logic, the way world is, and our understanding of it.

2002 208 pp., 8 line illus. $35.00



### FINITE AND INFINITE GOODS

**A Framework for Ethics**

**ROBERT MERRIHEW ADAMS**

“. . . Adams covers a tremendous amount of intellectual ground with such precision and rigor that this book will satisfy both theologians and philosophers alike....an indispensable text.”—*Religious Studies Review*.

“... a work that every Christian philosopher and theologian concerned with ethics simply must read.”—*Books & Culture*.

1999 (paper April 2002) 432 pp. paper $24.95 cloth $49.95

### KNOWLEDGE AND REALITY

**Selected Essays**

**COLIN MCGINN**

This book brings together a selection of Colin McGinn's philosophical essays from the 1970s to the 1990s, whose unifying theme is the relation between the mind and the world. The essays range over a set of prominent topics in contemporary philosophy, including the analysis of knowledge, the a priori, necessity, possible worlds, realism, mental representation, appearance and reality, and color.

1999 (paper 2002) 344 pp. paper $19.95 cloth $65.00

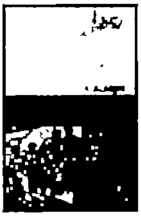
### INFINITE MINDS

**A Philosophical Cosmology**

**JOHN LESLIE**

The cosmos exists because of the ethical need for it. We, our universe, exist as thoughts in a divine mind that knows everything worth knowing. There could be infinitely many other universes in this mind. It may be hard to believe that the universe is as Leslie says it is—but it is also hard to resist his compelling ideas and arguments.

2002 261 pp. $35.00



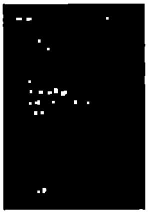
### EPICTETUS

**A Stoic and Socratic Guide to Life**

**A. A. LONG**

A. A. Long, a leading scholar of later ancient philosophy, gives the definitive presentation of the thought of Epictetus for a broad readership. Long's fresh and vivid translations of a selection of Epictetus's discourses show that his ideas are as valuable, relevant, and striking today as they were almost 2000 years ago.

2002 330 pp. $29.95



### IN NATURE'S INTERESTS?

**Interests, Animal Rights, and Environmental Ethics**

**GARY E. VARNER**

Varner offers a powerful response to what he calls the “two dogmas of environmental ethics”—the assumptions that animal rights philosophies and anthropocentric views are each antithetical to sound environmental policy.

“The book is a pleasure to read. The author is unduly influenced neither by radical nor conservative environmentalist camps. He has a mind of his own.”—*Environmental Ethics*

(*Environmental Ethics and Science Policy Series*)

1998 (paper 2002) 184 pp. paper $19.95 cloth $45.00

OXFORD

## ON VIRTUE ETHICS

**ROSALIND HURSTHOUSE**

"With this book virtue ethics finally comes of age. Hursthouse elegantly dispels the aura of unattractive high-mindedness that has clung to the approach ... she shows how the life of the virtuous is both possible and even enjoyable. This volume will effortlessly take its place as the defining exposition of the view."—Simon Blackburn, University of North Carolina

2000 (paper 2002) 288 pp paper $19.95

## KANT'S METAPHYSICS OF MORALS

**Interpretative Essays**

**Edited by MARK TIMMONS**

This is the only book devoted entirely to *The Metaphysics of Morals*. Seventeen essays by leading contemporary Kant scholars cover such topics as Kant's views on rights, punishment, contract, practical reasoning, revolution, freedom, virtue, legislation, happiness, moral judgement, love, respect, duties to oneself, and motivation.

May 2002 360 pp. paper $24.95 cloth $72.00

## LEGISLATING MORALITY

**Pluralism and Religious Identity in Lawmaking**

**LUCINDA PEACH**

The debate over religious lawmaking pits respect for religious pluralism against moral identity. Peach contends that both sides of the argument are fundamentally flawed and that neither has addressed the gender-based disparities of religious lawmaking. The book offers a pragmatic solution that respects religious pluralism, moral identity, and gender differences.

2002 232 pp. $35.00

## FREEDOM AND MORAL SENTIMENT

**Hume's Way of Naturalizing Responsibility**

**PAUL RUSSELL**

"...excellent creative scholarship."—Terence Penelhum, *Canadian Journal of Philosophy*.

"This book is a meticulous, wide-ranging reexamination of Hume's views on liberty, necessity and moral responsibility....Russell's account imbues Hume's texts with fresh significance and interest."—Ira Singer, *Ethics*.

1995 (paper 2002) 216 pp. paper $19.95 cloth $49.95

## LANGUAGE AND TIME

**QUENTIN SMITH**

"...[H]is arguments are clearly stated, forceful, and thought provoking....a *tour de force* that is required reading for those working in the philosophy of time."—L. Nathan Oaklander, University of Michigan at Flint

"Quentin Smith has produced a masterpiece...[that] will be regarded as the classic statement of the new tensed theory of time."—*The Philosophical Review*

1993 (paper May 2002) 288 pp. paper $24.95

Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556. In Canada, call 1-800-387-8020. Visit our website at **www.oup.com**.

OXFORD

## CLASSIC WORKS IN THE HISTORY OF IDEAS FROM THOEMMES PRESS

Thoemmes Press's high-quality, multi-volume editions bring together the major writings of important figures, works that exemplify key themes and debates, and the lesser-known and neglected works of the past—including modern classics previously out of print and indispensable reference works that are publishing firsts. In North America, Thoemmes Press books are distributed by the University of Chicago Press

For more information on these and other Thoemmes Press books,
visit their web site at www.thoemmes.com

### AMERICAN THEORIES OF POLYGENESIS
**Edited by Robert Bernasconi**

3808 pages, 7 Volumes 1-85506-947-4
Cloth $995.00

### LIVES OF MEN OF LETTERS AND SCIENCE WHO FLOURISHED IN THE TIME OF GEORGE III
**Lord Henry Brougham**

1866 pages, 2 Volumes 1-85506-948-2
Cloth $265.00

### WORKS OF SAMUEL CLARKE
**Samuel Clarke**

3254 pages, 4 Volumes 1-85506-968-7
Cloth $825.00

### THE POSITIVE PHILOSOPHY OF AUGUSTE COMTE
**Auguste Comte**
*Translated by Harriet Martineau*

1100 pages, 2 Volumes 1-85506-937-7
Cloth $265.00

### SYSTEM OF POSITIVE POLITY
**Auguste Comte**
*Translated by John Henry Bridges*, et al.

2496 pages, 4 Volumes 1-85506-936-9
Cloth $525.00

### THE ST. LOUIS HEGELIANS
**Edited by Michael DeArmey and James A. Good**

840 pages, 3 Volumes 1-85506-861-3
Cloth $295.00

THE UNIVERSITY OF CHICAGO PRESS
1427 East 60th Street, Chicago, IL 60637 • www.press.uchicago.edu

(Arv)

# THE JOURNAL OF PHILOSOPHY

02/4/03

VOLUME XCIX, NUMBER 8
AUGUST 2002

page
391 *The Harder Problem of Consciousness* Ned Block
426 *Carnap's Tolerance, Meaning, and Logical Pluralism* Greg Restall

444 NOTES AND NEWS

*Published by* **The Journal of Philosophy, Inc.**

# THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

*Editors*: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721. You may also visit our website at: www.journalofphilosophy.org

# THE JOURNAL OF PHILOSOPHY

2002

SUBSCRIPTIONS (12 issues)

| | |
|---|---|
| Individuals | $35.00 |
| Libraries and Institutions | $75.00 |
| Students, retired/unemployed philosophers | $20.00 |
| Postage outside the U.S. | $15.00 |

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Ephrata, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying More information on our terms may be obtained by writing to us.

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.
POSTMASTER: Send address changes to the *Journal of Philosophy* at MC 4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.
© Copyright 2002 by the Journal of Philosophy, Inc. ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY
VOLUME XCIX, NO. 8, AUGUST 2002

02\4\03

# THE HARDER PROBLEM OF CONSCIOUSNESS*

T. H. Huxley[1] famously said: "How it is that anything so remarkable as a state of consciousness comes about as a result of irritating nervous tissue, is just as unaccountable as the appearance of Djin when Aladdin rubbed his lamp" (*ibid.*, p. 19). We do not see how to explain a state of consciousness in terms of its neurological basis. This is the ***hard problem** of consciousness*.[2]

My aim here is to present another problem of consciousness. The ***harder problem***, as I shall call it, is more epistemological than the hard problem. A second difference: the hard problem could arise for

---

* I would like to thank David Barnett, Paul Boghossian, Tyler Burge, Alex Byrne, David Chalmers, Hartry Field, Jerry Fodor, Paul Horwich, Brian Loar, Tom Nagel, Georges Rey, Stephen Schiffer, and Stephen White for comments on earlier drafts. I am also grateful to Alex Byrne and Jaegwon Kim for reactions when an ancestor of this paper was delivered at a Central APA meeting in 1998. My thanks to the Colloquium on Language and Mind at New York University at which an earlier version of this paper was discussed in 2000, and especially to Tom Nagel as the chief inquisitor. I am also grateful to the audience at the 2001 meeting of Sociedad Filosofica Ibero Americana (SOFIA) and especially to my respondents, Brian McLaughlin and Martine Nida-Rümelin. And I would also like to thank my graduate class at NYU for their comments, especially Declan Smithies. In addition, I am grateful for discussion at a number of venues where earlier versions of this article were delivered, beginning with the Society for Philosophy and Psychology meeting, June 1997.

[1] *Lessons in Elementary Physiology* (New York: Macmillan, 1866); see Güven Güzeldere, "The Many Faces of Consciousness: A Field Guide," in Block, Owen Flanagan and Güzeldere, eds., *The Nature of Consciousness: Philosophical Debates* (Cambridge: MIT, 1997), pp. 1-67, footnote 6.

[2] See Thomas Nagel, "What Is It Like to Be a Bat?" *The Philosophical Review*, LXXXIII (1974): 435-50. Joseph Levine introduced the "explanatory gap" terminology (to be used later)—"Materialism and Qualia: The Explanatory Gap," *Pacific Philosophical Quarterly*, LXIV (1983): 354-61. David Chalmers and Galen Strawson distinguished between the hard problem and various "easy problems" of how consciousness functions—Chalmers, *The Conscious Mind* (New York: Oxford, 1996), pp. xxii-xxiii; Strawson, *Mental Reality* (Cambridge: MIT, 1994), pp. 93-96.

0022-362X/02/9908/391–425

someone who has no conception of another person, whereas the harder problem is tied closely to the problem of other minds. Finally, the harder problem reveals an epistemic tension or at least discomfort in our ordinary conception of consciousness which is not suggested by the hard problem, and so in one respect is harder. Perhaps the harder problem includes the hard problem and is best thought of as an epistemic add-on to it. Or perhaps they are in some other way facets of a single problem. Then my point is that this single problem breaks into two parts, one of which is more epistemic, involves other minds, and involves an epistemic tension.

## I. PRELIMINARIES

I believe that the major ontological disputes about the nature of consciousness rest on an opposition between two perspectives:

(1) *Deflationism about consciousness*, in which a priori or at least armchair analyses of consciousness (or at least armchair sufficient conditions) are given in nonphenomenal terms, most prominently in terms of representation, thought, or function.

(2) *Phenomenal realism*, which consists in the denial of deflationism plus the claim that consciousness is something real. Phenomenal realism is metaphysical realism about consciousness and thus allows the possibility that there may be facts about the distribution of consciousness which are not accessible to us even though the relevant functional, cognitive, and representational facts are accessible. Phenomenal realism is based on one's first-person grasp of consciousness. An opponent might prefer to call phenomenal realism "inflationism," but I reject the suggestion of something bloated.

In its most straightforward version, deflationism is a thesis of a priori conceptual analysis, most prominently analysis of mental terms in functional terms. As David Lewis,[3] a well-known deflationist noted, this view is the heir of logical behaviorism. Phenomenal realism rejects these armchair, philosophical-reductive analyses. But phenomenal realists have no brief against *scientific* reduction of consciousness. Of course, there is no sharp line here, and since the distinction is epistemic, one and the same metaphysical thesis could be held both as a philosophical-reductionist and as a scientific-reductionist thesis.[4]

---

[3] "An Argument for the Identity Theory," this JOURNAL, LXIII, 1 (January 6, 1966): 17-25.

[4] Deflationism with respect to truth is the view that the utility of the concept of truth can be explained disquotationally and that there can be no scientific reduction of truth—Paul Horwich, *Truth* (Cambridge: Blackwell, 1990; second edition, New York: Oxford, 1998); Hartry Field, "Deflationist Views of Meaning and Con-

I apologize for all the "isms" (deflationism, phenomenal realism, and one more to come), but they are unavoidable since my point here is that there is some tension between two of them. The tension is between phenomenal realism ("inflationism") and (scientific) naturalism, the epistemological perspective according to which the default view is that consciousness has a scientific nature—where this is taken to include the idea that conscious similarities have scientific natures. (A view on a given subject is the default if it is the only one for which background considerations give rational ground for tentative belief.) I argue for a conditional in which specifications of phenomenal realism and scientific naturalism (and a few other relatively uncontroversial items—including, notably, a rejection of a skeptical perspective) appear on the left-hand side. On the right-hand side, we have a specification of the epistemic tension which I mentioned. Deflationists who accept the argument may opt for modus tollens, giving them a reason to reject phenomenal realism. Phenomenal-realist naturalists may want to weaken their commitment to naturalism or to phenomenal realism. To put the point without explicit "isms": many of us are committed to the idea that consciousness is both real and can be assumed to have a scientific nature, but it turns out that these commitments do not fit together comfortably.

Modern phenomenal realism has often been strongly naturalistic (for example, Joseph Levine, Brian Loar, Colin McGinn, Christopher Peacocke, John Perry, Sydney Shoemaker, John Searle, and myself[5]). Daniel Dennett[6] has often accused phenomenal realists of closet dualism. Georges Rey[7] has argued that the concept of consciousness

---

tent," *Mind*, CIII (1994): 249-85. Deflationism with respect to consciousness in its most influential form is, confusingly, a kind of reductionism—albeit armchair reductionism rather than substantive scientific reductionism—and thus the terminology I am following can be misleading. I may have introduced this confusing terminology in my review of Daniel Dennett's *Consciousness Explained* in this JOURNAL, XC, 4 (April 1993): 181-93.

[5] Levine; Loar, "Phenomenal States," in Block, Flanagan, and Güzeldere, eds., pp. 597-616; McGinn, *The Problem of Consciousness* (Cambridge: Blackwell, 1991); Peacocke, *Sense and Content* (New York: Oxford, 1983), Perry, *Knowledge, Possibility and Consciousness* (Cambridge: MIT, 2001); Shoemaker, "The Inverted Spectrum," this JOURNAL, LXXIX, 7 (July 1982): 357-81; Searle, *The Rediscovery of the Mind* (Cambridge: MIT, 1992)

[6] *Consciousness Explained* (Boston: Little Brown, 1991).

[7] "A Reason for Doubting the Existence of Consciousness," in R. Davidson, G. Schwartz, D. Shapiro, eds., *Consciousness and Self-Regulation, Volume 3* (New York: Plenum, 1983), pp. 1-39. In previous publications—"On a Confusion about a Function of Consciousness," *The Behavioral and Brain Sciences*, XVIII, 2 (1995): 227-47—I have argued that Rey's alleged incoherence derives from his failure to distinguish between phenomenal consciousness and other forms of consciousness

is incoherent. The upshot here is that there is a grain of truth in these accusations.

Before I go on, I must make a terminological comment. Imagine two persons both of whom are in pain, but only one of whom is introspecting his pain state and is, in that sense, conscious of it. One could say that only one of the two has a *conscious* pain. This is *not* the sense of 'conscious' used here. In the sense of 'conscious' used here, just in virtue of *having* pain, *both* have conscious states. To avoid verbal disputes, we could call the sense of 'consciousness' used here *phenomenality*. Pains are intrinsically phenomenal and *in that sense* are intrinsically conscious. In that sense—but not in some other senses—there cannot be an unconscious pain.

My plan here is this: first, I shall briefly characterize the hard problem, mainly in order to distinguish it from the harder problem. I shall argue that the hard problem can be dissolved, only to reappear in a somewhat different form, but that in this different form we can see a glimmer of hope for how a solution might one day be found. I shall then move on to the harder problem, its significance and a comparison between the hard and harder problems. I shall conclude with some reflections on what options there are for the naturalistic-phenomenal realist.

## II MIND-BODY IDENTITY AND THE APPARENT DISSOLUTION OF THE HARD PROBLEM

The hard problem is one of explaining why the neural basis of a phenomenal quality is the neural basis of *that* phenomenal quality rather than another phenomenal quality or no phenomenal quality at all. In other terms, there is an explanatory gap between the neural basis of a phenomenal quality and the phenomenal quality itself. Suppose (to replace the neurologically ridiculous example of c-fiber stimulation, which is often used by philosophers, with a view proposed as a theory of visual experience by Francis Crick and Christof Koch[8]) that cortico-thalamic oscillation (of a certain sort) is the neural basis of an experience with phenomenal quality $Q$. Now, there is a simple (over-simple) physicalist dissolution to the hard problem which is based on mind-body identity: phenomenal quality $Q$ = cortico thalamic oscillation (of a certain sort). Here is a statement of the solution:

---

(what I call *access consciousness* and *reflexive consciousness*). The incoherence that is the subject of this article, by contrast, is an incoherence in phenomenal consciousness itself.

[8] "Towards a Neurobiological Theory of Consciousness," *Seminars in the Neurosciences*, II (1990): 263-75.

> The hard problem is illusory. One might as well ask why $H_2O$ is the chemical basis of water rather than gasoline or nothing at all. Just as water *is* its chemical basis, so $Q$ just is its neural basis (cortico-thalamic oscillation), and that shows the original question is wrongheaded.

I think there is something right about this answer, but it is nonetheless unsatisfactory. What is right about it is that, if $Q$ = cortico-thalamic oscillation, that identity itself, like all genuine identities, is inexplicable.[9] What is wrong about it is that we are in a completely different epistemic position with respect to such a mind-body identity claim than we are with respect to 'water $=H_2O$'. The claim that $Q$ is identical to cortico-thalamic oscillation is just as puzzling—maybe more puzzling—than the claim that the physical basis of $Q$ is cortico-thalamic oscillation. We have no idea how it could be that one property could be identical both to $Q$ and cortico-thalamic oscillation. How could one property be both subjective and objective? Although no one can explain an identity, we can remove puzzlement by explaining how an identity can be true, most obviously, how it is that the two concepts involved can pick out the same thing. This is what we need in the case of subjective/objective identities, such as the putative identity that $Q$ = cortico-thalamic oscillation.

## III. HOW TO APPROACH THE HARD PROBLEM

The standard arguments against physicalism (most recently by Frank Jackson, Saul Kripke, and David Chalmers[10]) make it difficult to understand how mind-body identity could be true, so explaining how it could be true requires undermining those arguments. I shall not attempt such a large task here, especially since the role of the present discussion of the hard problem is mainly to contrast it with the harder problem to come. So I shall limit my efforts in this direction to a brief discussion of Jackson's famous "knowledge" argument. I discuss this argument not because I think it is the most challenging argument against mind-body identity, but rather because it motivates an apparatus that gives us some insight into what makes the hard problem hard. Jackson imagined a neuroscientist of the distant future (Mary) who is raised in a black and white room and who knows everything physical and functional that there is to know about color and the experience of it. But when she steps outside the room for the first time, she learns what it is like to see red. Jackson argued that since the physical and functional facts do not encompass the new fact that Mary learns, dualism is true.

---

[9] We can reasonably wonder how it is that Mark Twain and Samuel Clemens married women with the same name, lived in the same city, and so on. But we cannot reasonably wonder how it is that Mark Twain *is* Samuel Clemens. This point is made in my "Reductionism," *Encyclopedia of Bioethics* (New York: Macmillan, 1978), pp. 1419-24. See also Block and Robert Stalnaker, "Conceptual Analysis and the Explanatory Gap," *The Philosophical Review*, CVIII (January 1999): 1-46; and David Papineau, "Physicalism, Consciousness, and the Antipathetic Fallacy," *Australasian Journal of Philosophy*, LXXI (1993): 169-83. For a statement of a contrary view, see Chalmers.

[10] Jackson, "What Mary Didn't Know," and Kripke, "The Identity Thesis," in Block, Flanagan, and Güzeldere, eds , pp. 567-70, and pp. 445-50; and Chalmers.

The key to what is wrong with Jackson's argument (and to removing one kind of puzzlement about how a subjective property could be identical to an objective property) is the concept/property distinction.[11] Any account of this distinction as it applies to phenomenal concepts is bound to be controversial. I shall use one such account without defending it, but nothing in what follows here will be based on this account.

The expressions 'this sudden involuntary muscle contraction' and 'this [experience] thing in my leg' are two expressions that pick out the cramp I am now having in my leg. (These are versions of examples from Loar.) In 'this [experience] thing in my leg', attention to an experience of the cramp functions so as to pick out the referent, the cramp. (That is the meaning of the bracket notation. The 'this' in 'this [experience] thing in my leg' refers to the thing in my leg, not the experience.) The first way of thinking about the cramp is an objective concept of the cramp. The second is a subjective concept of the same thing—subjective in that there is a phenomenal mode of access to the thing picked out. Just as we can have both objective and subjective concepts of a cramp, we can also have objective and subjective concepts of a cramp *feeling*. Assuming physicalism, we could have an objective neurological concept of a cramp feeling, for example, 'the phased locked 40 Hz oscillation that is occurring now'. And we could have a subjective concept of the same thing, 'this [experience] feeling'. Importantly, the same experience type could be part of—though function differently—in *both* subjective concepts, the subjective concept of the cramp and the subjective concept of the cramp feeling. Further, we could have both a subjective and objective concept of a single color. And we could have both a subjective and an objective concept of the experience of that color, and the same

[11] The articles in Block, Flanagan, and Güzeldere, eds., by Paul Churchland ("Knowing Qualia: A Reply to Jackson," pp. 571-77), Loar, William Lycan ("Consciousness as Internal Monitoring," pp 775-71), and Robert van Gulick ("Understanding the Phenomenal Mind: Are We All Just Armadillos? Part One, Phenomenal Knowledge and Explanatory Gaps," pp 559-66) all take something like this line; as does Scott Sturgeon, "The Epistemic View of Subjectivity," this JOURNAL, XCI, 5 (May 1994): 221-35; and Perry.

experience or mental image could function—albeit differently—in the two subjective concepts, one of the color, the other of the experience of the color.

Deflationists will not like this apparatus, but they should be interested in the upshot since it may be of use to them in rejecting the phenomenal realism in the antecedent of the conditional for which I am arguing here.

Concepts in the sense used here are mental representations. For our purposes, we may as well suppose a system of representation that includes both quasilinguistic elements as well as phenomenal elements, such as experiences or mental images. Stretching terminology, we could call it a language of thought.[12]

In these terms, then, we can remove one type of puzzlement that is connected with the hard problem as follows: there is no problem about how a subjective property can be identical to an objective property. Subjectivity and objectivity are better seen as properties of *concepts* rather than of *properties*. The claim that an objective property is identical to a subjective property would be more revealingly expressed as the claim that an objective concept of a property picks out the same property as a subjective concept of that property. So we can substitute a dualism of concepts for a dualism of properties.

The same distinction helps us to solve the Mary problem. In the room, Mary knew about the subjective experience of red via the objective concept *cortico-thalamic oscillation*. On leaving the room, she acquires a subjective concept *this [mental image] phenomenal property* of the same subjective experience. In learning what it is like to see red, she does not learn a new fact. She knew about that fact in the room under an objective concept, and she learns a new concept of that very fact. One can acquire new knowledge about old facts by acquiring new concepts of those facts. New knowledge acquired in this way does not show that there are any facts beyond the physical facts. Of course, it does require that there are concepts that are not physical*istic* concepts, but that is not a form of dualism. (For present purposes, we can think of physicalistic concepts as concepts couched in the vocabulary of physics. A physicalist can allow nonphysicalistic vocabulary, for example, the vocabulary of economics. Of course, physicalists say that everything is physical, including vocabulary. But the vocabulary of economics can be physical in that sense without being physicalististic in the sense of couched in the vocabulary of physics.)

---

[12] Note that my account of subjective concepts allows for subjective concepts of many more colors or pitches than we can recognize, and thus my account differs from accounts of phenomenal concepts as *recognitional* concepts (such as that of Loar). On my view, one can have a phenomenal concept without being able to reidentify the same experience again. For arguments that experience outruns recognition, see Sean Kelly, "Demonstrative Concepts and Experience," *The Philosophical Review*, cx, 3 (2001): 397-420.

Where are we? The hard problem in one form was: How can an objective property be identical to a subjective property? We now have a dissolution of one aspect of the problem, appealing to the fact that objectivity and subjectivity are best seen as properties of concepts. But that is no help in getting a sense of what *sorts* of objective concepts and subjective concepts could pick out the same property, and so it brings us no closer to actually getting such concepts. As Nagel (*op. cit.*) noted, we have no idea how there could be causal chains from an objective concept and a subjective concept leading back to the same phenomenon in the world. We are in something like the position of pre-Einsteinians who had no way of understanding how a concept of mass and a concept of energy could pick out the same thing.

## IV. PRELIMINARIES BEFORE INTRODUCING THE HARDER PROBLEM

*(A) Naturalism.* Naturalism is the view that it is a default that consciousness has a scientific nature (and that similarities in consciousness have scientific natures). I shall assume that the relevant sciences include physics, chemistry, biology, computational theory, and parts of psychology that do not explicitly involve consciousness. (The point of the last condition is to avoid the trivialization of naturalism that would result if we allowed the scientific nature of consciousness to be...consciousness.) I shall lump these sciences together under the heading 'physical', thinking of naturalism as the view that it is a default that consciousness is physical (and that similarities in consciousness are physical). So naturalism = default physicalism, and is thus a partly epistemic thesis. Naturalism in my sense recognizes that, although the indirect evidence for physicalism is impressive, there is little direct evidence for it. My naturalist is not a "die-hard" naturalist, but rather one who takes physicalism as a default, a default that can be challenged. My rationale for defining 'naturalism' in this way is that this version of the doctrine is plausible, widely held, and leads to the epistemic tension that I am expositing. Some other doctrines that could be called 'naturalism' do not, but this one does. I think that my naturalism is close to what Perry calls "antecedent physicalism" (*op. cit.*).

*(B) Functionalism.* Functionalism and physicalism are usually considered competing theories of mind. For present purposes, however, the phenomenal realism/deflationism distinction is more important, and this distinction cross-cuts the distinction between functionalism and physicalism. In the terms used above, one type of functionalism

is deflationist, the other phenomenal realist. The latter is *psychofunctionalism*, the identification of phenomenality with a role property specified in terms of a psychological or neuropsychological theory.[13] At the beginning, I pointed to the somewhat vague distinction between philosophical and scientific reduction. Deflationist functionalism is a philosophical-reductionist view whereas phenomenal-realist psychofunctionalism is a scientific-reductionist view.

I will be making use of the notion of a superficial functional isomorph, a creature that is functionally isomorphic to us with respect to those causal relations among mental states, inputs, and outputs that are specified by "folk psychology." Those who are skeptical about these notions should note that the point here is that a nexus of standard views leads to a tension. This conceptual apparatus may be part of what should be rejected.[14]

As I mentioned at the outset, I am arguing for a conditional. On the left side of the conditional are phenomenal realism and naturalism (plus conceptual apparatus of the sort just mentioned). My current point is that I am including psychofunctionalism in the class of phenomenal-realist naturalist theories. Thus, one kind of functionalism—the deflationist variety—is excluded by the antecedent of my conditional, and another—the phenomenal-realist variety—is in the class of open options.

*(C) Antiskeptical perspective.* In what follows, I shall be adopting a point of view that sets skepticism aside. "*Undoubtedly*, humans are conscious and rocks and laptops are not." (Further, *bats* are undoubtedly conscious.) Of course, the antiskeptical point of view I shall be adopting is the one appropriate to a naturalist phenomenal-realist. Notably, from the naturalist-phenomenal realist perspective, the concept of a functional isomorph of us with no consciousness is not incoherent and the claim of bare possibility of such a zombie—so long as it is not alleged to be us—is not a form of skepticism.

*(D) Multiple realization/multiple constitution.* Putnam, Jerry Fodor, and Block and Fodor[15] argued that, if functionalism about the mind

---

[13] See my "Troubles with Functionalism," in C. W. Savage, ed., *Minnesota Studies in the Philosophy of Science, Volume IX* (Minneapolis: Minnesota UP, 1978), pp. 261-325.

[14] Those who would like to see more on functionalism should consult any of the standard reference works such as the *Routledge Encyclopedia of Philosophy*. Or see http://www.nyu.edu/gsas/dept/philo/faculty/block/papers/functionalism.html.

[15] Putnam, "Psychological Predicates," later titled "The Nature of Mental States," in W. H. Capitan and D. D. Merrill, eds., *Art, Mind, and Religion* (Pittsburgh: University Press, 1967), pp. 37-48; Fodor, "Materialism," in *Psychological Explanation* (New York: Random House, 1968), chapter 3, pp. 90-120; Block and Fodor, "What Psychological States Are Not," *The Philosophical Review*, LXXXI (1972): 159-81.

is true, physicalism is false. The line of argument assumes that functional organizations are often—maybe even always—multiply realizable. The state of adding 2 cannot be identical to an electronic state if a nonelectronic device (for example, a brain) can add 2.

This "multiple-realizability" argument has become controversial lately,[16] for reasons that I cannot go into here. The argument I shall be giving is a version of the traditional multiple-realizability argument (albeit an epistemic version), so I had better say a bit about what a realization is. One of the many notions of realization that would do for present purposes is the following. A functional state is a kind of second-order property, one that consists in having certain first-order properties that have certain causes and effects.[17] For example, dormitivity in one sense of the term is the property a pill has of having some (first-order) property that causes sleep. Provocativity is the property of having some (first-order) property or other that makes bulls angry. We can speak of the first-order property of being a barbiturate as being one realizer of dormitivity, or of red as being one realizer of provocativity.[18]

If we understand realization, we can define constitution in terms of it. Suppose that mental state *M* has a functional role that is realized by neural state *N*. Then *N* constitutes *M*—relative to *M*'s playing the *M*-role. The point of the last condition is that ersatz *M*—a state functionally like *M* but missing something essential to *M* as phenomenality is to pain—would also have the *M*-role, but *N* would not constitute ersatz *M* merely in virtue of constituting *M*. So the *M*-role can be multiply realized even if *M* is not multiply constituted.

---

[16] See Jaegwon Kim, "Multiple Realization and the Metaphysics of Reduction," *Philosophy and Phenomenological Research*, LII (1992): 1-26; and *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation* (Cambridge: MIT, 1998).

[17] The restriction to first order properties is unnecessary. See my "Can the Mind Change the World?" in George Boolos, ed , *Meaning and Method: Essays in Honor of Hilary Putnam* (New York: Cambridge, 1990), pp. 137-70.

[18] An alternative notion of realization appeals to the notions of supervenience and explanation. The realized property supervenes on the realizer and the realizer explains the presence of the realized property. Possessing the realizer is one way in which a thing can possess the realized property. See Ernest Lepore and Barry Loewer, "Mind Matters," this JOURNAL, LXXXIV, 11 (November 1987)· 630–42; and Lenny Clapp, "Disjunctive Properties: Multiple Realizations," this JOURNAL, XCVIII, 3 (March 2001): 111-36.

Dormitivity in the sense mentioned is a second-order property, the property of having some property that causes sleep. But one could also define dormitivity as a first-order property, the property of causing sleep. That is, on this different definition, *F* is dormitive just in case *F* causes sleep. But if we want to ascribe dormitivity to *pills*, we shall have to use the second-order sense What it is for a pill to be dormitive is to have some property or other that causes sleep. Similarly, if we want a notion of functional property that applies to properties, the first-order variant will do. But if we want to ascribe those properties to people, we need second-order properties.

There is an obvious obscurity in what counts as *multiple* realization (or constitution). We can agree that neural property *X* is distinct from neural property *Y* and that both realize a single functional property without agreeing on whether *X* and *Y* are variants of a single property or two substantially different properties, so we shall not agree on whether there is genuinely multiple realization. And even if we agree that *X* and *Y* are substantially different, we may still not agree on whether the functional property is multiply realized since we may not agree on whether there is a single disjunctive realization. These issues will be discussed further in section x.

## V. INTRODUCING THE HARDER PROBLEM

My strategy will be to start with the epistemic possibility of multiple realization and use it to argue for the epistemic possibility of multiple constitution of mentality. I shall then argue that the epistemic possibility of multiple constitution of phenomenal properties is problematic. I shall use a science-fiction example of a creature who is functionally the same as us but physically different. Those who hate science fiction should note that the same issue arises—in more complicated forms—with respect to real creatures, such as the octopus, which differ from us both physically and functionally.

> (C1): We have no reason to believe that there is any deep physical property in common to all and only the possible realizations of our superficial functional organization.

Moreover—and this goes beyond what is needed for (C1), but it does make (C1) more vivid—we have no reason to believe that we cannot find or *make* a merely superficial isomorph of ourselves. By "merely superficial isomorph," I mean an isomorph with respect to folk psychology and whatever is logically or nomologically entailed by folk psychological isomorphism, but that is all. For example, the fact that pains cause us to moan (in circumstances that we have some appreciation of but no one has ever precisely stated) is known to common sense, but the fact that just-noticeable differences in stimuli increase with increasing intensity of the stimuli (the Weber-Fechner law) is not. So the merely superficial isomorph would be governed by the former but not necessarily the latter. The television series *Star Trek: The Next Generation* includes an episode ("The Measure of a Man," broadcast 2/26/89) in which there is a trial in which it is decided whether a human-like android, Commander Data, may legally be turned off and taken apart by someone who does not know whether he can put the parts together again. (The technology that allowed the android to be built has been lost.[19]) Let us take Commander Data to be a merely superficial isomorph of us (ignoring his superior reasoning and inferior emotions). Then (C1) can be taken to be that we have no reason to believe that Commander Data is not nomologically or otherwise metaphysically possible. Note that I am not making so strong a claim as made in Block and Fodor (*op. cit.*)—that there is empirical reason to suppose that our functional organization is multiply realizable—but only that we have no reason to doubt it.

The strategy of the argument, you recall, is to move from the epistemic possibility of multiple realization to the epistemic possibility of multiple constitution. (C1) is the epistemic possibility of multiple realization.

> (C2): Superficial functional equivalence to us is a defeasible reason for attributing consciousness; that is, superficial functional equivalence to us provides a reason for thinking a being is conscious, but that reason can be disarmed or unmasked, its evidential value cancelled.

(C2) consists of two claims, that superficial functional equivalence to us is a reason for attributing consciousness and that that reason is defeasible. The first claim is obvious enough. I am not claiming that the warrant is a priori, just that there is warrant. I doubt that there will be disagreement with such a minimal claim.

What is controversial about (C2) is that the reason is claimed to be defeasible. Certainly, deflationary functionalists will deny the defeasibility. Of course, even deflationary functionalists would allow that *evidence* for thinking something is functionally equivalent to us can be defeated. For example, that something emits English sounds is a reason to attribute consciousness, but if we find the sound is recorded, the epistemic value of the evidence is cancelled. But (C2) does not merely say that functional or behavioral *evidence* for consciousness can be defeated. (C2) says that, even if we *know* that something is functionally equivalent to us, there are things we can find out that cancel the reason we have to ascribe consciousness (without challenging our knowledge of the functional equivalence). A creature's consciousness can be unmasked without unmasking its functional equivalence to us.

Here is a case in which the epistemic value of functional isomorphism is cancelled: the case involves a *partial physical* overlap between the functional isomorph and humans. Suppose that there are real neurophysiological differences of kind—not just complexity—between our conscious processes and our unconscious—that is, nonphenomenal—processes. Nonphenomenal neural process include, for example, those which regulate body temperature, blood pressure, heart rate, and sugar in the blood—brain processes that can operate in people in irreversible vegetative coma. Suppose (*but only temporarily*—this assumption will be dispensed with later) that we find out that *all* of the merely superfical isomorph's brain states are ones that—in us—are the neural bases *only of phenomenally unconscious states.* For example, the neural basis of the functional analog of pain in the merely superficial isomorph is the neural state that regulates the pituitary gland in us. This would not *prove* that the isomorph is not phenomenally conscious (since the contexts of the neural realizers are different), but it would cancel or at least weaken the force of the reason for attributing consciousness provided by its functional isomorphism to us.

---

[19] There is a synopsis by Timothy Lynch at http://www.ugcs.caltech.edu/st-tng/episodes/135.html.

The role of this case is to motivate a further refining of our characterization of Commander Data and to justify (C2) by exhibiting the epistemic role of a defeater.

Let us narrow down Commander Data's physical specification to rule out the cases just mentioned as defeaters for attribution of consciousness to him. Here is a first shot:

(i) Commander Data is a superficial isomorph of us.

(ii) Commander Data is a merely superficial isomorph. So we have no reason to suppose there are any shared physical properties between our conscious states and Commander Data's functional analogs of them that could be the physical basis of any phenomenal overlap between the two, since we have no reason to think that such shared properties are required by the superficial overlap. Further, one could imagine this discussion taking place at a stage of science where we could have rational ground for believing that there are no shared nondisjunctive physical properties (or more generally scientific properties) that could be the physical basis of a phenomenal overlap. Note that no stipulation can rule out *certain* shared physical properties—for example, the disjunctive property of having the physical realizer of the functional role of one of our conscious states or Commander Data's functional analog of it.

(iii) The physical realizers of Commander Data's functional analogs of conscious states do not overlap with any of our brain mechanisms in any properties that we do not also share with inorganic entities that are uncontroversially mindless, for example, toasters. So we can share properties with Commander Data like having molecules. But none of the realizers of Commander Data's analogs of conscious states are the same as realizers of, for example, our states that regulate our blood sugar—since these are organic.

(iv) Commander Data does not have any part which itself is a functional isomorph of us and whose activities are crucial to maintaining the functional organization of the whole.[20]

The point of the last two conditions is to specify that Commander Data has a realization that cannot be seen to defeat the attribution of consciousness to him either a priori or on the basis of a theory of *human* consciousness. (For example, the last condition rules out a "homunculi-headed" realization.) It would help if I could think of all the realizations that have these kinds of significance. If you tell me about one I have not thought of, I shall add a condition to rule it out.

*Objection:* we are entitled to reason from same effects to same causes. Since our phenomenal states play a role in causing our behavior, we can infer that the functionally identical behavioral states of Commander Data are produced in the same way. The only other option, the objection goes on, is epiphenomenalism.

*Reply:* consider two computationally identical computers, one that works via electronic mechanisms, the other that works via hydraulic mechanisms. (Suppose that the fluid in one does the same job that the electricity does in the other.) We are not entitled to infer from the causal efficacy of the fluid in the hydraulic machine that the electrical machine also has fluid. One need not be an epiphenomenalist to take seriously the hypothesis that there are alternative realizations of the functional roles of our phenomenal states that are phenomenally blank.[21]

We might suppose, just to get an example on the table, that the physical basis of Commander Data's brain is to be found in etched silicon chips rather than the organic carbon basis of our brains.

The reader could be forgiven for wondering at this point whether I have not assembled stipulations that close off the question of Commander Data's consciousness. Naturalism includes the doctrine that it is the default that a conscious overlap requires a physical basis, and it may seem that I have in effect stipulated that Commander Data does not have any physical commonality with us that could be the basis of any shared phenomenality. The objection ignores the option of a shared *disjunctive* basis and certain other shared bases to be discussed below.

---

[20] Following Putnam. This stipulation needs further refinement, which it would be digressive to try to provide here.

[21] See Shoemaker; he makes assumptions that would dictate that Commander Data overlaps with us in the most general phenomenal property, *having phenomenality*—in virtue of his functional likeness to us. But in virtue of his lack of physical overlap to us, there are no shared phenomenal states other than phenomenality itself. So on Shoemaker's view, phenomenality is a functional state, but more specific phenomenal states have a partly physical nature.

(C3): Fundamentally different physical realization from us per se is not a ground of rational belief in lack of consciousness. So the fact that Commander Data's control mechanisms are fundamentally different is not a ground of rational belief that he has no phenomenal states. Note that I do not say that finding out that Commander Data has a silicon-based brain is not a *reason* for regarding him as lacking consciousness. Rather, I say that the reason falls below the epistemic level of a ground for rational belief.

(C4): We have no conception of a ground of rational belief to the effect that a realization of our superficial functional organization that is physically fundamentally different along the lines I have specified for Commander Data is or is not conscious. To use a term suggested by Martine Nida-Rümelin in commenting on this paper, Commander Data's consciousness is meta-inaccessible. Not only do we lack a ground of belief, but we lack a conception of any ground of belief. This meta-inaccessibility is a premise rather than a lemma or a conclusion because the line of thought I have been presenting leads up to it without anything that I am happy to think of as an argument for it. My hope is that this way of leading up to it will allow the reader to see it as obvious.

We can see the rationale for meta-inaccessibility by considering Searle's chinese room argument.[22] He famously argued that even if we are computational creatures, we are not either sentient or sapient merely in virtue of that computational organization. In reply to his critics,[23] he says repeatedly that a machine that shares our computational organization and is therefore behaviorally and functionally equivalent to us—and therefore passes the Turing Test—need not be an intentional system (or a conscious being). What would make it an intentional system—and for Searle, intentionality is engendered by and requires consciousness—is not the functional organization but rather the way that functional organization is implemented in the biology of the organism. But, to take an example that Searle uses, how would we know whether something made out of beer cans is sentient or sapient? He says: "It is *an empirical question* whether any given machine [that shares our superficial functional organization] has causal powers equivalent to the brain" (*ibid.*, p. 452). "I think it is evident that all sorts of substances in the world, like water pipes and toilet paper, are going to lack those powers, but that is ***an empirical claim*** on my part. On my account it is a ***testable empirical claim*** whether in repairing a damaged brain," we could duplicate these causal powers (*ibid.*, p. 453). "I offer no a priori proof that a system of integrated circuit chips could not have intentionality. That is, as I say repeatedly, ***an empirical question.*** What I do argue is that in order to produce intentionality the system would have to duplicate the causal powers of the brain and that simply instantiating a formal program would not be sufficient for that" (*ibid.*, p. 453; emphasis added).

---

[22] "Minds, Brains, and Programs," *Behavioral and Brain Sciences*, III (1980): 417-24; see also *The Rediscovery of the Mind*.

[23] "Author's Response," *The Behavioral and Brain Sciences*, III (1980): 450-57.

I do not deny that one day the question of whether a creature like Commander Data is phenomenally conscious may ***become*** a testable empirical question. But it is obvious that we do not ***now*** have any conception of how it could be tested. Searle has suggested (in conversation) that the question is an empirical one in that, if I were the device, I would know from the first-person point of view if I were conscious. But even if we were to accept such a counterfactual, we could not take it as showing that the claim is testable or empirical in any ordinary sense of the term.

Although I am tweaking Searle's flamboyant way of putting the point, my naturalist phenomenal-realist view is not that different from his. I agree that whether physically different realizations of human functional organization are conscious is not an a priori matter and could be said to depend on whether their brains have "equivalent causal powers" to ours—in the sense of having the power to be the physical basis of conscious states. (But I do not agree with Searle's view that the neural bases of conscious states "cause" the conscious states in any normal sense of 'cause'.) I agree with him that consciousness is a matter of the biology of the organism, not (just) its information processing. The issue that I am raising here for naturalist phenomenal-realism affects my view as much as his.

I am not denying that we might some day come to have the conception we now do not have. (So I am not claiming—as McGinn does (*op. cit.*)—that this knowledge can be known now to be beyond our ken.) I am merely saying that, at this point, we have no idea of evidence that would ground rational belief, even a hypothetical or speculative conception. Of course, those who meet Commander Data will reasonably be sure that he is conscious. But finding out that he is not ***human*** cancels that ground of rational belief.

Perhaps we shall discover the nature of human consciousness and find that it applies to other creatures. For example, the nature of human consciousness may involve certain kinds of oscillatory processes that can apply to silicon creatures as well. But the problem I am raising will arise in connection with realizations of our functional organization that lack those oscillatory processes. The root of the epistemic problem is that the example of a conscious creature on which the science of consciousness is inevitably based is us (where "us" can be construed to include nonhuman creatures that are neurologically similar to humans). But how can science based on us generalize to creatures that do not share our physical properties? It would seem that a form of physicalism that could embrace other creatures would have to be based at least in part on them in the first place, but that cannot be done unless we already know whether they are conscious.

I have left a number of aspects of the story unspecified. What was the aim of Commander Data's designer? Just how fine grained is the functional isomorphism? I hope that for each of my readers there is *some* way of filling in the story for which that reader will agree that we have no ground of rational belief either way.

I keep using the phrase 'ground of rational belief'. What does it mean? I take this to be an epistemic level that is stronger than 'reason for believing' and weaker than 'rational certainty'. I take it that a ground of rational belief that *p* allows knowledge that *p* but mere reason for believing *p* does not.

## VI. DISJUNCTIVISM AND THE EPISTEMIC PROBLEM

I now move to the conditional that I advertised earlier. Let us start by supposing, but only temporarily, that physicalism requires a deep (nonsuperficial) unitary (nondisjunctive) scientific (physical) property shared by all and only conscious beings. This version of physicalism seems at first glance to be incompatible with Commander Data's being conscious, and the corresponding version of naturalism (which says that physicalism is the default) seems at first glance to be epistemically incompatible with phenomenal realism. That is, naturalism says the default is that Commander Data is not conscious, but phenomenal realism says that the issue is open in the sense of no rational ground for belief either way. This is a first pass at saying what the harder problem is.

If this strong kind of physicalism really is incompatible with Commander Data's being conscious, we might wonder whether the reasons we have for believing physicalism will support this weight. I shall pursue a weaker version of physicalism (and corresponding version of naturalism) which does not rule out consciousness having a physical basis that is disjunctive according to the standards of physics. As we shall see, however, the stronger version of physicalism is *not* actually incompatible with Commander Data's being conscious, and the difference between the stronger and weaker versions makes no important difference with respect to our epistemic situation concerning Commander Data's consciousness.

Disjunctivism is a form of physicalism that allows that consciousness is a physical state that is disjunctive by the standards of physics. As applied to the current issue, disjunctivism allows that if Commander Data is conscious, the shared phenomenality is constituted by the property of having Commander Data's electronic realization of our shared functional state or our electrochemical realization.

In footnote 16, I mentioned Jaegwon Kim's critique of the multiple-realizability argument against physicalism. He argues that, if mental property $M$ is nomically equivalent to a heterogeneous disjunction $N$, we should regard $M$ as non-nomic and non-"real" because $N$ is. He argues that, if human thought can be realized by very different physical mechanisms from, say, Martian or robot thought, then the real sciences of thought will be the sciences of the separate realizations of it. To call them all 'thought' is simply to apply a superficial verbal concept to all of them, but the laws of human thought will be different from the laws of Martian thought. The real kinds are not at the level of the application of verbal concepts.[24]

Even those who are sympathetic to this picture of thought must make an exception for consciousness (in the sense, as always in this article, of phenomenality). We can be happy with the view that there is a science of human thought and another science of machine thought, but no science of thought per se. But we should not be happy with the idea that there is a science of human phenomenality, another of machine phenomenality, and so on. For since the overlap of these phenomenalities, *phenomenality*, is something real and not merely nominal as in the case of thought, it must have a scientific basis. If a phenomenal property is nomically coextensive with a heterogeneous neural disjunction, it would not be at all obvious that we should conclude that the phenomenal property is non-nomic and non-"real" because the disjunction is. The phenomenal-realist naturalist point of view would be more friendly to the opposite, that the disjunction is nomic and "real" because the phenomenal property is.

The real problem with disjunctivism is that whether it is true or not, we could have no good reason to believe it. To see this, we shall have to have a brief incursion into the epistemology of reductive identity.

---

[24] See my replies to Kim, "Anti-reductionism Slaps Back," *Mind, Causation, World: Philosophical Perspectives*, XI (1997): 107-33; and "Do Causal Powers Drain Away?" forthcoming in *Philosophy and Phenomenological Research*, with a reply by Kim.

## VII THE EPISTEMOLOGY OF THEORETICAL IDENTITY

Why do we think that water = $H_2O$, temperature = mean molecular kinetic energy and freezing = lattice formation?[25] The answer begins with the fact that water, temperature, freezing, and other magnitudes form a family of causally interrelated "macro"-properties. This family corresponds to a family of "micro"-properties: $H_2O$, mean molecular kinetic energy, formation of a lattice of $H_2O$ molecules. And the causal relations among the macroproperties can be explained if we suppose the following relations between the families: that water = $H_2O$, temperature = mean molecular kinetic energy, and freezing = lattice formation. For example, as water is cooled, it contracts until about 4 degrees (F) above freezing, at which point it expands. Why? Why does ice float on water? Here is a sketch of the explanations: the oxygen atom in the $H_2O$ molecule has two pairs of unmated electrons, which attract the hydrogen atoms on other $H_2O$ molecules. Temperature = mean molecular kinetic energy. When the temperature (namely, kinetic energy) is high, the kinetic energy of the molecules is high enough to break these hydrogen bonds, but as the kinetic energy of the molecules decreases, each oxygen atom tends to attract two hydrogen atoms on the ends of two other $H_2O$ molecules. When this process is complete, the result is a lattice in which each oxygen atom is attached to four hydrogen atoms. Ice is this lattice and freezing is the formation of such a lattice. Because of the geometry of the bonds, the lattice has an open, less dense structure than amorphously structured $H_2O$ (namely, liquid water)—which is why ice (solid water) floats on liquid water. The lattice forms slowly, beginning about 4 degrees above freezing. (The exact temperature can be calculated on the basis of the numerical values of the kinetic energies needed to break or prevent the bonds.) The formation of large open lattice elements is what accounts for the expansion of water on the way to freezing. (Water contracts in the earlier cooling because decreasing kinetic energy allows more bonding, and until the bonding reaches a stage in which there are full lattice elements, the effect of the increased bonding is make the water more densely packed.)

Suppose we reject the assumption that temperature is identical to mean molecular kinetic energy, in favor of the assumption that temperature is merely correlated with mean molecular kinetic energy. And suppose we reject the claim that freezing is lattice formation, in favor of a correlation thesis; and likewise for water/ $H_2O$. Then we would have an explanation for how something that is *correlated* with decreasing temperature causes something that is *correlated* with frozen water to float on something *correlated* with liquid water, which is not all that we want. Further, if we assume identities, we can explain why certain macroproperties are spatiotemporally coincident with certain microproperties. The reason to think that the identities are true is that assuming them gives us explanations that we would not otherwise have and does not deprive us of explanations that we already have or raise explanatory puzzles that would not otherwise arise. The idea is not that our reason for thinking these identities are true is that it would be nice if they were true. Rather, it is that assuming that they are true yields the *most explanatory overall picture*. In other words, the epistemology of theoretical identity is just a special case of inference to the best explanation.

---

[25] The temperature identity is oversimplified, applying in this form only to gases. Paul Churchland raises doubts about whether there is a more abstract identity—*Matter and Consciousness* (Cambridge: MIT, 1984) I think those doubts are deflected in Simon Blackburn's "Losing Your Mind: Physics, Identity and Folk Burglar Prevention," in *Essays in Quasi-Realism* (New York: Oxford, 1993), chapter 13.

Some suppose that substance identities such as 'water = $H_2O$' are on a different footing from "property" identities, and that substance identities can be established on purely spatiotemporal grounds.[26] But deciding that water and $H_2O$ *are spatiotemporally coincident* is part of the same package as having decided that they are one and the same. For example, the air above a glass of water buzzes with bits of water in constant exchange with water in the atmosphere, a fact that we can acknowledge only if we are willing to suppose that those $H_2O$ molecules *are* bits of water. The claim that water is $H_2O$ and that water and $H_2O$ are spatiotemporally coincident stand or fall together as parts of one explanatory package. And once we conclude that the substance liquid water = amorphous $H_2O$ and that the substance frozen water = lattice-structured $H_2O$, we would be hard pressed to deny that freezing = lattice formation, since the difference between liquid and frozen water is that the former has an amorphous structure and the latter a lattice structure. Substance identities and property identities often form a single explanatory package.

## VIII. BACK TO DISJUNCTIVISM

With the epistemology of identity in place, we can now ask whether there could be an argument from inference to the best explanation to the conclusion that consciousness is a heterogeneous physical disjunction, the disjunction of our realization of the consciousness role and Commander Data's corresponding realization. Of course without a prior decision as to whether Commander Data's states are actually conscious, there could be no such argument. Putting this point aside, let us suppose, temporarily, that Commander Data is conscious. Even so, the prospects for an argument from inference to the best explanation to the identity of a phenomenal property with a disjunctive physical property are dubious. Molten glass hardens into an amorphous solid-like substance. (If there are absolutely no impurities, fast continuous cooling of water can make it harden without lattice formation in a similar manner.) We could give a disjunctive explanation of solid-like formation that included both freezing and this kind of continuous hardening. And if we preferred that disjunctive explanation to two distinct explanations, we would regard the hardening of glass as a kind of freezing and glass as a kind of solid. We do not regard amorphous hardening as freezing. But we do not take the disjunctive explanation seriously and so we regard glass as (strictly speaking) a super-cooled liquid rather than a solid. We have an important though vague notion of "fundamentally different" that governs our willingness to regard some differences in realization as variants of the same basic type and others as fundamentally different. Without going into the difficult issue of what it is that makes two realizations fundamentally different, we can stipulate that our physical realizations of our conscious states are fundamentally different from Commander Data's physical realizations of his analogs of these states. And this fundamental difference precludes an argument to the best explanation of the sort I have been considering.

---

[26] Kim gave a paper at Columbia University in December 1999 making this suggestion, and Tim Maudlin argued that all theoretical identities are established on spatiotemporal grounds when I gave this paper at Rutgers.

One of the epistemic factors relevant to our acceptance of a theoretical identity is what questions it rules out. The question of why it is that water is correlated with $H_2O$ or why it is that heat is correlated with molecular kinetic energy are bad questions, and they are ruled out by the identity claims that water $= H_2O$ and heat $=$ molecular kinetic energy. But the question of why we overlap phenomenally with Commander Data in this respect rather than that respect or some respect rather than none do not seem illegitimate. It is hard to think of a reason why we should accept a physicalistic view that dictates that they are illegitimate—disjunctivism—rather than opt for a nonphysicalist view that holds out some hope for an answer.

*Objection*: you say identities cannot be explained, but then you also say that we can have no reason to accept a disjunctive physicalistic identity because it is not explanatory.

*Reply*: identities cannot be explained, but they can contribute to explanations of other things. My point about the epistemology of identity is that it is only because of the explanatory power of identities that we accept them and the disjunctive identity countenanced by disjunctivism does not pass muster.

Disjunctivism is one way of making naturalism compatible with Commander Data's being conscious, but there are others. One is the view that consciousness is as a matter of empirical fact identical to the superficial functional organization that we share with Commander Data. We might call this view *superficialism* (with apologies to Rey who has used this term for a somewhat different doctrine). Recall that the phenomenal realist/deflationist distinction is an epistemic one, so any ontological view could in principle be held as having either epistemic status. Superficialism is the *phenomenal-realist* claim that consciousness is identical to the superficial functional organization that we share with Commander Data—as distinct from the deflationist version of this claim mentioned earlier.

Note that superficialism says consciousness is a role property, not a property that *fills* or realizes that role. A role property is a kind of dispositional property. Now, there is no problem about dispositions being caused: spraying my bicycle lock with liquid nitrogen causes it to become fragile. So if pain is a superficial functional state, we can perhaps make use of that identification to explain the occurrence of pain in neural terms. Whether dispositions are causes—as would be required by this identity—is a more difficult issue that I shall bypass. (Does a disposition to say ouch cause one to say ouch?)

The difficulty I want to raise is that, even if identifying pain with a superficial functional role does license explanations of the *superficial* causes and effects of being in pain, the identification cannot in the same way license explanations of the *nonsuperficial* causes and effects of being in pain. Suppose, for example, that psychologists discover that pain raises the perceived pitch of sounds. Even if we take the thesis that pain is a disposition to say ouch to help us to explain why pain causes saying ouch, it will not explain the change in pitch. The epistemic difficulty I am pointing to is that there is no good reason why the causal relations *known to common sense* ought to be explained differently from the ones not known to common sense. So the identification raises an explanatory puzzle that would not otherwise arise, and that puts an epistemic roadblock in the way of the identification. This is perhaps not a conclusive difficulty with the proposal, but it does put the burden of proof on the advocate of the identification to come up with explanatory advantages so weighty as to rule out the explanatory disadvantage just mentioned.[27]

Of course, this objection will not apply to the phenomenal realist identification of consciousness with its *total* functional role as opposed to its superficial functional role. Since the physiology of Commander Data's states differs from ours, their total functional roles will differ as well. So this would be a chauvinist proposal that would beg the question against Commander Data's consciousness.

Nida-Rümelin objected that there are a vast number of properties, maybe infinitely many, that are entailed nomologically or logically by the superficial functional equivalence, and each of these is both shared with Data and is a candidate for the nature of consciousness. Certainly, a full treatment would attempt to categorize these properties and assess their candidacy. Some—for example, possessing complex inputs and outputs—can be eliminated because they are also shared with mindless computers. Of course, there may be others that are not so easily dismissed.

### IX THE UPSHOT

I said earlier that it seemed at first glance that a form of physicalism which required that consciousness be constituted by a unitary physical property dictated that Commander Data is not conscious. We can now see that, at second glance, this is not the case. Even if we preclude a disjunctive physical basis to the phenomenal overlap between us and Commander Data (assuming that there is such an overlap), still the physicalist could allow that Commander Data is conscious *on superficialist grounds*. And even if we reject superficialism, there are other potential meta-inaccessible physical bases of a phenomenal overlap between us and Commander Data.

The upshot is that physicalism in neither the stronger nor weaker versions mentioned above rules out Commander Data's being conscious. But *the only epistemically viable naturalist or physicalist hypothesis*—the only naturalist or physicalist hypothesis we have a conception of a reason for accepting—is a deep unitary physical or otherwise scientific property in common to all and only conscious beings, a naturalistic basis that Commander Data does not share. So for the physicalist, Commander Data's consciousness is not epistemically viable.

Thus, our knowledge of physicalism is *doubly* problematic: we have no conception of a ground of rational belief that Commander Data is or is not conscious, and we have no way of moving from a conclusion that Commander Data is conscious to any consequence for the truth of physicalism. And this holds despite the fact that physicalism is our default view. *Physicalism is the default and also inaccessible and meta-inaccessible.* This is part of the harder problem. A second part follows.

---

[27] I am grateful to Chalmers for pressing me for a better treatment of this issue.

But first I shall discuss the question of whether the epistemic tension itself is a good reason to conclude that Commander Data is not conscious. The short version of my answer is that, while the epistemic tension is a bad consequence of our phenomenal-realist view that it is an open question whether Commander Data is conscious, it is not the kind of bad consequence that justifies us in concluding that he is not conscious. I shall justify this claim.

*Objection:* You say disjunctivism is epistemically defective, but is it not also metaphysically defective? How could a unitary phenomenal property be identical to a physical property that is nonunitary?

*Reply:* there is no logical flaw in disjunctivism. If a unitary phenomenal property is identical to a nonunitary physical property, then one property is both unitary from the mental point of view and nonunitary from the physical point of view. We are willing to allow that unitary properties of economics, sociology, and meteorology are nonunitary from the physical point of view. Why should we not include mentality, too?

There is more to this issue than I can go into here.[28] Of course, there are views that are worthy of being called 'naturalism' that dictate that disjunctivism is metaphysically defective. But they are not the 'naturalism' that I am talking about. The naturalist I am talking about, you will recall, is also a phenomenal realist. And being a phenomenal realist, this naturalist keeps the question open of whether creatures that are heterogeneous from a physical point of view nonetheless overlap phenomemenally. If you like, this is a naturalistic concession to phenomenal realism.

*Objection:* silicon machinery of the sort we are familiar with is manifestly not conscious. The only reason we could have to suppose that Commander Data's brain supported consciousness would be to find some kind of physical similarity to the states that we know underlie human consciousness, and that possibility has been ruled out by stipulation. Moreover, we can explain away our tendency to think of Commander Data as conscious as natural but unjustified anthropomorphizing.

*Reply:* naturalism and phenomenal realism do not dictate that Commander Data is not conscious or that the issue of his consciousness is not open. Recall that disjunctivism and superficialism are metaphysically (though not epistemically) viable. Further, naturalism gives us no evidence against or reason to doubt the truth of either disjunctivism or superficialism. Hence naturalism (and physicalism) give us no good reason to doubt the the consciousness of Commander Data. Imagine arguing at Commander Data's trial that he is a zombie (or that there is no matter of fact as to whether he is conscious), while conceding that his zombiehood is not even *probabilified* by naturalism unless we set aside disjunctivism and superficialism, options on which he may be conscious. And imagine conceding that we are setting these options aside not because we have any evidence against them but because we cannot conceive of any way in which they may be known. He could reasonably say (or to be neutral, produce the noise): "Your lack of a conception of how to find out whether I am conscious is no argument that I am a zombie; I similarly lack a conception of how to find out whether you are conscious." In any case, phenomenal realism is a form of (Putnamian) metaphysical realism, so the phenomenal realist cannot suppose that our ignorance, even necessary ignorance, is a reason to suppose that Commander Data is not conscious or that there is no matter of fact as to whether he is.

---

[28] See my "Anti-reductionism Slaps Back," for more on this topic.

Phenomenal realists regard consciousness as a real property that does not have logical ties to behavior. That is what allows phenomenal realists to take seriously the application of consciousness or protoconsciousness to elementary particles. But if anything *might* be conscious, why take the consciousness of anything other than humans seriously? One answer can be seen by considering what happens if one asks Commander Data whether red is closer to purple than blue is to yellow. Answering such questions requires, in us, a complex multidimensional phenomenal space—in part captured by the color solid—with phenomenal properties at many levels of abstractness (cf. Loar). Commander Data's functional equivalence to us guarantees that he has an internal space that is functionally equivalent to our phenomenal space. But anyone who grasps our phenomenal space from the first-person point of view has to take seriously the possibility that an isomorphic space in another being is grasped by him from a similar first-person perspective. Talking of our "functional equivalence" to Commander Data tends to mask the fact that we are like him in a complex structure or set of structures. If one thinks of the functional similarity as limited to saying 'Ouch' when you stick a pin in him, it is easy to miss the positive phenomenal realist rationale for regarding Commander Data's consciousness as an open question.

Thus, the phenomenal realist and the deflationist converge on not closing off the possibility that Commander Data is conscious.

*Objection* (made by many critics): Why should the mere epistemic possibility of a bad consequence of physicalism threaten physicalism? No one thinks that the mere epistemic possibility of an object that has mass traveling faster than light threatens relativity theory. If relativity is true, nothing can travel faster than light. Similarly, if physicalism is true, there is no conscious Commander Data.

*Reply*: first, relativity theory gives us reason to believe that it is impossible for anything to travel faster than light. But naturalism and phenomenal realism do not give us reason to believe that there can be no Commander Data or that it is impossible that Commander Data is conscious. The *die-hard* physicalist may take the fact that Commander Data leads to unpalatable consequences as a reason to deny the existence of the consciousness of Commander Data. But the form of naturalism that leads to the epistemic problem is a more liberal and more plausible form. It is not die-hard physicalism, but rather physicalism that recognizes that we lack much in the way of direct evidence for the view. Second, the trouble that Commander Data leads to is not of the sort that gives us rational ground for denying the existence of a conscious Commander Data. Disjunctivism is not metaphysically suspect but only epistemically suspect: we have no conception of how to know whether it is true or not. Our lack of knowledge is no argument against the consciousness of Commander Data.

*Objection*: Brian McLaughlin[29] has argued that I am mischaracterizing the epistemic role of functional similarity in our reasoning about other minds. The role of functional similarity is in providing evidence that others are like us in intrinsic physical respects, and that is the ground for our belief in other minds. In the case of Commander Data, that evidential force is cancelled when we find out what Commander Data's real constitution is. McLaughlin notes that we are happy to ascribe consciousness to babies even though they are functionally very different from us because we have independent evidence that they share the relevant intrinsic physical properties with us. The same applies, though less forcefully, to other mammals—for example, rabbits. He asks us to compare a human baby with a functionally equivalent robot baby. The robot baby's functional equivalence to the real baby gives us little reason to believe that the robot baby is conscious. Similarly, for the comparison between a real rabbit and a robot rabbit. Moving closer to home, consider a paralytic with Alzheimers: little functional similarity to us, but we are nonetheless confident, on the basis of an inference from similarity in intrinsic physical properties, that the senile paralytic has sensory consciousness. The upshot, he says, is that material constitution and structure trumps function in our attribution of consciousness to others. And so, if we become convinced that Commander Data is unlike us in the relevant intrinsic physical respects, we should conclude that he is not conscious.

---

[29] In a response at SOFIA, 2001.

*Reply*: first, Commander Data shares with us *disjunctivist* and *superficialist material constitution and structure*, and so no conclusion can be drawn about the consciousness of Commander Data, even if McLaughlin is right about material constitution and structure trumping function. Nothing in McLaughlin's argument supplies a reason to believe that disjunctivism or superficialism are false. (Recall that I have argued that these views are *epistemically* defective, not that they are false.) He says that the relevant physical properties are "intrinsic" but if that is supposed to preclude disjunctivism or superficialism, we are owed an argument. Second, I do agree with McLaughlin that a substantial element of our belief in other consciousnesses depends on an inference to a common material basis. It would be a mistake to conclude, however, that this inference provides the entire basis for our attribution of other consciousnesses. Our justification is an inference from like effects to like causes. Even if we find out that the causes of behavioral similarity are not alike in material constitution and structure, it remains open that the common cause is a similarity in *consciousness itself* and that consciousness itself has a disjunctive or superficial material basis or no material basis. (Recall that naturalism is committed to physicalism as a default, but a default can be overridden.)

Third, function is not so easy to disentangle from material constitution and structure, at least epistemically speaking. The opponent process theory of color vision originated in the nineteenth century from common-sense observations of color vision, such as the fact that afterimages are of the complementary color to the stimulus and that there are colors that seem, for example, both red and blue (purple) or red and yellow (orange), but no color that seems both red and green or both blue and yellow. The basic two-stage picture of how color vision works (stage 1: three receptor types; stage 2: two opponent channels) was discovered before the relevant physiology on the basis of behavioral data. To the extent that Commander Data behaves as we do, there is a rationale for supposing that the machinery of Commander Data's color vision shares an abstract structure with ours that goes beyond the color solid.

The first of the epistemic difficulties on the right-hand side of our conditional is that physicalism is the default, but also inaccessible and meta-inaccessible. We are now ready to state the second epistemic difficulty. Let us introduce a notion of the *subjective default* view which we have rational ground for believing on the basis of background information—but only ignoring escape hatches—such as disjunctivism and superficialism—which we have no evidence against but which are themselves inaccessible and meta-inaccessible. Then the second epistemic difficulty is that of holding *both that it is an open question whether Commander Data is conscious and that it is the subjective default view that he is not.* These two epistemic difficulties constitute the harder problem.

Before I go on to consider further objections, let me briefly contrast the point of this paper with Nagel's famous "bat" paper (*op. cit.*). His emphasis was on the functional differences between us and bats, creatures which share the mammalian physical basis of sensation. My example, however, is one of a functionally identical creature, the focus being on the upshot of physical differences between us and that creature.

The issue of the application of our phenomenal concepts to exotic creatures is often mentioned in the literature, but assimilated to the hard problem (the "explanatory gap"). (I am guilty, too. That was the background assumption of the discussion of "universal psychology" in my "Troubles with Functionalism" (*op. cit.*).) For example, Levine notes that we lack a principled basis for attributing consciousness to creatures which are physically very different from us. He says: "I submit that we lack a principled basis precisely because we do not have an explanation for the presence of conscious experience even in ourselves" (*op. cit.*, p. 79). Later, he says: "Consider again the problem of attributing qualia to other creatures, those that do not share our physical organization. I take it that there is a very real puzzle whether such creatures have qualia like ours or even any at all. How much of our physicofunctional architecture must be shared before we have similarity or identity of experience? This problem, I argued above, is a direct manifestation of the explanatory gap" (*op. cit.*, p. 89).

## X. MORE OBJECTIONS

One can divide objections into those which require clarification of the thesis and those which challenge the thesis as clarified. The objections considered so far are more in the former category while those below are more in the latter.

*(A) Objections from indeterminacy. Objection:* the issue of whether Commander Data is conscious is just a matter of vagueness or inde-

terminacy in the word 'conscious'. If we reject property dualism, then the issue of whether Commander Data is conscious depends on extrapolating a concept of consciousness grounded in our physical constitution to other physical constitutions. If those other physical constitutions are sufficiently different from ours as is stipulated for Commander Data, then the matter is indeterminate and so a decision has to be made. Similarly, in extending the concept 'wood' to an alien form of life, we might find that it resembles what we have already called 'wood' in certain ways but not others and a decision will have to be made.[30]

*Reply*: no phenomenal realist—physicalist or not—should accept the assumption that the decision whether to attribute consciousness to Commander Data is a decision about whether to extrapolate from our physical constitution to his, and so the objection presupposes phenomenal irrealism. According to the phenomenal realist, Commander Data may share our conscious qualities even if he does not resemble us physically.

I do not want to give the impression that phenomenal realism is incompatible with indeterminacy about consciousness. For example, perhaps a fish is a borderline case of consciousness. Similarly, Commander Data might be a borderline case of consciousness and therefore indeterminate. On the phenomenal realist view of consciousness, it is an open question whether Commander Data is (a) conscious, (b) not conscious, (c) a borderline case. But there is no reason to think that Commander Data *must* be a borderline case. From the phenomenal-realist point of view, epistemic considerations *alone* do not show metaphysical indeterminacy.

*(B) Closure of epistemic properties.* In a response to this article, Nida-Rümelin[31] gave a formalization of the argument which involved a principle of closure of epistemic properties such as being open or being meta-inaccessible. (Brendan Neufeld made a similar point.) For example, she supposes that part of the argument goes something like this: supposing physicalism requires a deep unitary property in common to conscious creatures, if Commander Data is conscious, then physicalism is false; Commander Data's consciousness is meta-inaccessible; so the falsity of physicalism is meta-inaccessible.

One can easily see that the form of argument is fallacious. If Plum did it, then it is false that the butler did it. But if it is inaccessible whether Plum did it, it does not follow that it is inaccessible whether or not the butler did it. We might find evidence against the butler which has nothing to do with Plum. The application of the point to the argument that Nida-Rümelin attributes to me is that, even if Commander Data's consciousness is inaccessible, we might have some independent reason to believe physicalism is false. I explicitly noted (and did in the earlier version) that I think the standard arguments against physicalism do not work.

---

[30] Field and Papineau have pressed such views in commenting on an earlier version of this article, though they are not responsible for the type of example.

[31] At SOFIA, 2001.

Here is a standard problem with closure.[52] Consider a meta-inaccessible claim, *I*, and an accessible claim, *A*. The conjunction *I* & *A* is meta-inaccessible, but a consequence of it, *A*, is not. So meta-inaccessibility is not transmitted over entailment. Briefly and metaphorically: fallacies of the sort mentioned arise with respect to an epistemic property that applies to a whole even if only one of its parts has that property. The whole can then entail a different part that does not have that epistemic property. I doubt that my argument has that form, but if someone can show that it does, that will undermine it.

(C) *Objections concerning empirical evidence. Objection*: suppose my brain is hooked up to Commander Data's and I have the experience of seeing through his eyes. Is that not evidence that he has phenomenal consciousness?

*Reply*: maybe it is evidence, but it does not get up to the level of a rational ground for believing. Perhaps if I share a brain in that way with a zombie, I can see through the zombie's eyes, because whatever is missing in the zombie brain is made up for by mine.

*Objection*: suppose we discover what we take to be laws of consciousness in humans and discover that they apply to Commander Data. That is, we find that the laws that govern human consciousness also govern the functional analog of consciousness in Commander Data. Does that not get up to the level of rational ground for believing that Commander Data is conscious?[53]

*Reply*: since Commander Data's brain works via different principles from ours, it is *guaranteed that his states will not be governed by all of the same laws* as the functionally equivalent states in us. Two computers that are computationally equivalent but physically different are inevitably different in all sorts of physical features of their operation, for example, how long they take to compute various functions, and their failure characteristics—such as how they react to humidity or magnetic fields. The most that can be claimed is that the state that is the functional analog of human consciousness in Commander Data obeys *some* of the laws that our conscious states obey. The problem is: Are the laws that Commander Data does *not* share with us laws of consciousness or laws of his *different physical realizer*? Without an understanding of the scientific nature of consciousness, how are we supposed to know? A zombie might share some laws of consciousness, but not enough or not the right ones for consciousness. So long as Commander Data does not share *all* the laws of our conscious states, there will be room for rational doubt as to whether the laws that he does share with us are decisive. Indeed, if we knew whether Commander Data was conscious or not, we could use that fact to help us in deciding which laws were laws of consciousness and which were laws of the realization. But as this point suggests, the issue of whether Commander Data is conscious is of a piece with the epistemic problem of whether a given law is a law of consciousness or a law of one of the realizers of its functional role.

---

[52] See my discussion of the tacking paradox in "Anti-Reductionism Slaps Back."

[53] I am grateful to Barry Smith for getting me to take this objection more seriously.

An example will be useful to clarify this point. All human sensory systems obey a power function, an exponential function relating stimulus intensity to subjective intensity as judged by subjects' reports. That is, subjective intensity = stimulus intensity raised to a certain exponent, a different exponent for different modalities. For example, perceived brightness is proportional to energy output in the visible spectrum raised to a certain exponent. This applies even to outré parameters of subjective judgments, such as how full the mouth feels as a function of volume of wadges of paper stuck in the mouth or labor pains as a function of size of contractions. Should we see the question of whether Commander Data's sensations follow the power law as a litmus test for whether he has conscious experiences? No doubt the power law taps some neural feature. Is that neural feature essential or accidental to the nature of consciousness? Roger Shepard[34] has argued that the power law form would be expected in any naturally evolved creature. But that leaves open the possibility of artificial creatures or evolutionary singularities (subject to unusual selection pressures) whose sensations (or "sensations") do not obey the power law. The question whether this is a law of consciousness or a law of the human realization of consciousness that need not be shared by a conscious Commander Data is of a piece with the question of whether creatures like Commander Data (who, let us suppose, do not obey the law) are conscious. We cannot settle one without the other, and the epistemic problem I am raising applies equally to both.

---

[34] In his unpublished William James Lectures.

*(D) Skepticism and the problem of other minds.* Recall that I am arguing for a conditional. On the left are naturalism, phenomenal realism, and the denial of skepticism. There is a superficial resemblance between the harder problem and the problem of other minds. But the problem of other minds is a form of skepticism. The nonskeptic has no doubt that *humans* are (sometimes) conscious, but when we find out that Commander Data is *not human,* denying skepticism does not help.

What is it about being human that justifies rejecting skepticism? It is not part of my project here to attempt an answer, but I have to say *something* to avoid the suspicion that our rationale for regarding other humans as conscious or rocks as not conscious might apply equally to Commander Data.

Elliot Sober's "Evolution and the Problem of Other Minds"[35] argues plausibly that our rationale for attributing mental states to other humans is a type of "common-cause" reasoning. But such common-cause reasoning is vulnerable to evidence against a common cause, for example, evidence for lack of genealogical relatedness or evidence for different scientific bases for the similarity of behavior that is exhibited. Thus the rationale for attributing mentality to humans does not fully apply to Commander Data.

Stephen White[36] raises the skeptical worry of how we know that creatures whose brains are like ours in terms of principles of operation but not in DNA are conscious. But this worry may have a *scientific* answer that would be satisfying to the nonskeptic. We might arrive at a partial understanding of the mechanisms of human consciousness which is sufficient to assure us that a creature that shared those mechanisms with us is just as conscious as we are even if its DNA is different. For example, we might discover a way to genetically engineer a virus that replaced the DNA in the cells of living creatures. And we might find that when we do this for adult humans such as ourselves, there are no noticeable effects on our consciousness. Or we might come to have something of a grip on why cortico-thalamic oscillation of a certain sort is the neural basis of human consciousness and also satisfy ourselves that many changes in DNA in adults do not change cortico-thalamic oscillation. By contrast, the harder problem may remain even if we accept the dictates of nonskeptical science.

---

[35] This JOURNAL, XCVII, 7 (July 2000): 365-86.

[36] "Curse of the Qualia," *Synthese,* LXVIII (1983): 333-68; reprinted in Block, Flanagan, and Güzeldere, eds., pp. 695-718. See also Shoemaker, pp. 653-54.

## XI SUPERVENIENCE AND MIND-BODY IDENTITY

Much of the recent discussion of physicalism in the philosophy of mind has centered on supervenience of consciousness on the brain rather than on good old-fashioned mind-body identity. Chalmers recommends this orientation, saying "I find that discussions framed in terms of identity generally throw more confusion than light onto the key issues, and often allow the central difficulties to be evaded. By contrast, supervenience seems to provide an ideal framework within which key issues can be addressed" (*op. cit.*, p. xvii).

But the harder problem depends on the puzzling nature of multiple physical constitution of consciousness, a problem that does not naturally arise from the perspective that Chalmers recommends. Supervenience prohibits any mental difference without a physical difference, but multiple constitution is a physical difference without a mental difference. Of course nothing prevents us from stating the issue in supervenience terms. In those terms, it is the problem of how a unitary phenomenal property can have a nonunitary (heterogeneously disjunctive) supervenience base. But there is no reason why this should be puzzling from the supervenience point of view. Heterogeneous supervenience bases of unitary properties—for example, adding—are common. What makes it puzzling is the thought that a phenomenal overlap between physically different creatures *ought* to have a unitary physical basis. That puzzle can be appreciated from the point of view of old-fashioned mind-body identity—which says that a phenomenal overlap is a physical property. (No one would identify adding with a physical (for example, microphysical) property—it is obviously functional.) But it is not puzzling from the supervenience point of view.

## XII. THE HARD AND THE HARDER

Are the hard and harder problems really different problems? The hard problem is: Why is the scientific basis of a phenomenal property the scientific basis of that property rather than another or rather than a nonphenomenal property? The question behind the harder problem could be put so as to emphasize the similarity: Why should physically different creatures overlap phenomenally in one way rather than another or not at all? This way of putting it makes it plausible that the harder problem includes or presupposes the hard problem. In any case, the harder problem is more narrowly epistemic than the hard problem. The hard problem could arise for someone who has no conception of another person, whereas the harder problem is closely tied to the problem of other minds. Finally, the harder problem involves an epistemic tension not involved in the hard problem. My claim is that the "harder problem" differs from the "hard problem" in these ways independently of whether we choose to see them as distinct problems or as part of a single problem.

Is the harder problem harder than the hard problem? If the harder problem is the hard problem plus something else, then it is trivially harder. As indicated above, the harder problem has an epistemic dimension not found in the hard problem, so they are to that extent incomparable, but the epistemic difficulty involved in the harder problem makes it harder in one way.

Both the hard and harder problems depend on what we cannot *now* conceive. Even the epistemic difficulty may be temporary, unlike the epistemic difficulty of the concept of *the gold mountain that no one will ever have evidence of.* Perhaps we will come to understand the nature of human consciousness, and in so doing, develop an objective theory of consciousness that applies to all creatues, independently of physical constitution. That is, perhaps the concepts developed in a solution to the hard problem will one day solve the harder problem, though I think our relation to this question is the same as to the harder problem itself—namely, we have no conception of how to find an answer.

## XIII. WHAT TO DO?

Naturalism dictates that physicalism is the default, but also inaccessible and meta-inaccessible; and in the "subjective" sense mentioned earlier, it is the default that Commander Data is not conscious, but at the same time phenomenal realists regard his consciousness as an open issue. What to do? To begin, one could simply live with these difficulties. They are not paradoxes. Physicalism is the default and at the same time meta-inaccessible. It is the subjective default that androids like Commander Data are not conscious, but it is an open question whether they are. Consciousness is a singularity—perhaps one of its singular properties is thrusting us into epistemic tension. Another option would be to reject or restrict the assumption of naturalism or of phenomenal realism. One way to debase naturalism slightly would be to take the problem itself as a reason to believe the disjunctivist or superficialist form of naturalism. Those who prefer to weaken phenomenal realism can do so without adopting one of the deflationist views mentioned at the outset: functionalism, representationism, and cognitivism. One way to restrict phenomenal realism is to adopt what Shoemaker (*op. cit*) calls the "Frege-Schlick" view that comparisons of phenomenal character are only meaningful within the stages of a single person and not between individuals. Another proposal is slightly weaker than the Frege-Schlick view in allowing only interpersonal comparisons across naturalistically similar persons. That is, although comparisons of phenomenal character among subjects who share a physical (or other naturalistic) basis of that phenomenal character make sense, comparisons outside that class are nonfactual; or else a significant group of them are false. That is, Commander Data either has no consciousness or there is no matter of fact about his consciousness.

Naturalistic phenomenal-realism is not an unproblematic position. We cannot comfortably suppose both that consciousness is real and that it has a scientific nature. I have not argued here for one or another way out, but have been concerned only with laying out the problem.

NED BLOCK

New York University

# CARNAP'S TOLERANCE, MEANING, AND LOGICAL PLURALISM*

With J. C. Beall, I[1] have argued for and defended a pluralism about logical consequence. We take it that the notion of logical consequence is irreducibly plural in its application. That is, we take it that there are at least two distinct relations of logical consequence—and not simply two distinct relations in intension, but two distinct relations in extension. We take it that there are arguments that are valid according to one logic, and invalid according to another, and that there is no further fact of the matter as to whether the argument is *really* valid. For the sake of this discussion, let me be quite specific. I shall take two particular argument forms. First, the argument form I shall call *explosion*:

From $A$ and $\neg A$ to infer $B$

This argument form is valid classically, and invalid relevantly. Similarly, the argument form of *disjunctive syllogism*:

From $A \vee B$ and $\neg A$ to infer $B$

is valid classically, and invalid relevantly.[2] As a pluralist about logical consequence, I take it that there is no further fact of the matter as to whether explosion, or disjunctive syllogism are *really* valid. For me, that question makes no more sense than to ask whether a function on the real line is *really* smooth, without saying more about the notion of smoothness. A function might be smooth in the sense of *continuity* without being smooth in the sense of *differentiability*. The same goes

---

* Thanks to Peter Simons for asking the question "Isn't your pluralism just Carnap's tolerance?" I hope that I have managed to explain why it is not. Many thanks to J. C. Beall for discussion on the topics raised in this paper. Although we have written extensively on pluralism together and I have profited greatly from working with him, the errors and infelicities in this paper are all mine and not his Thanks also to Martin Allen, Robert Anderson, Nuel Belnap, Otávio Bueno, Daniel Nolan, Koji Tanaka, and audiences at the Center for Philosophy of Science at the University of Pittsburgh, and the Philosophy Departments at Macquarie University and Universities of Melbourne, Toronto, and Queensland for constructive comments on this paper. This research has been supported by the Australian Research Council, through Large Grant A00000348, and by Koji Tanaka who relieved me of some of my teaching duties for Semester 2, 2000.

[1] "Logical Pluralism," *Australasian Journal of Philosophy*, LXXVIII (2000): 475-93.

[2] For an extended discussion of this issue, see Nuel D. Belnap and J. Michael Dunn, "Entailment and the Disjunctive Syllogism," in F. Fløistad and G. H. von Wright, eds., *Philosophy of Language / Philosophical Logic* (The Hague: Martinus Nijhoff, 1981), pp. 337-66.

0022-362X/02/9908/426–43

with logical consequence. Logical consequence is an important *pretheoretic* notion which must be made precise if it is to have definite application in all cases. This much is not a controversial thesis, and it dates back at least to Alfred Tarski.[3] Tarski acknowledged that he was doing more than this: he was aware that in proposing a *formalization* of logical consequence, he was choosing between alternatives.

> The concept of *logical consequence* is one of those whose introduction into the field of strict formal investigation was not a matter of arbitrary decision on the part of this or that investigator; in defining this concept, efforts were made to adhere to the common usage of the language of everyday life. But these efforts have been confronted with the difficulties which usually present themselves in such cases. With respect to the clarity of its content the common concept of consequence is in no way superior to other concepts of everyday language. Its extension is not sharply bounded and its usage fluctuates. Any attempt to bring into harmony all possible vague, sometimes contradictory, tendencies which are connected with the use of this concept, is certainly doomed to failure. We must reconcile ourselves from the start to the fact that every precise definition of this concept will show arbitrary features to a greater or less degree (*ibid.*, p. 409).

Our account of logical pluralism agrees with Tarski. Logical consequence is a pretheoretic notion which may be made precise in a number of different ways. In some ways, this is a claim that the concept of logical consequence is ambiguous. Our logical pluralism is more than this, however: we have argued that the core notion of logical consequence is not *arbitrarily* ambiguous, but rather, it has a simple structure: a conclusion is a logical consequence of some premises if and only if that conclusion is true in every *circumstance* in which the premises are true. Different disambiguations of this notion arise from taking different extensions of the term "circumstance."

I shall not say very much more to elaborate the different notions of logical consequence I endorse. Beall and I have explored these notions at greater length elsewhere.[4] Suffice to say that the relation I shall here call *classical logical consequence* agrees with classical propositional logic at the level of conjunction, disjunction, and negation. I also take it that, given a classically valid argument, it is never the case that the premise is true and the conclusion is not. That is, I *endorse* classical reasoning, in the sense that I think that I shall never step from truth to untruth in taking the step from premises to a conclusion in a classically valid argument.[5]

---

[3] "On the Concept of Logical Consequence," in *Logic, Semantics, Metamathematics: Papers from 1923 to 1938*, J. H. Woodger, trans. (New York: Oxford, 1956), pp. 409-20.

[4] Cf. Beall and Restall; and my "Negation in Relevant Logics: How I Stopped Worrying and Learned to Love the Routley Star," in Dov Gabbay and Heinrich Wansing, eds., *What Is Negation* (Boston: Kluwer, 1999), pp. 53-76.

The relation I shall call *relevant logical consequence* agrees with *first-degree entailment* as defined in the tradition arising out of Alan Ross Anderson and Nuel Belnap's[6] work on entailment. First-degree entailment rejects the inference of explosion and its dual (from $A$ to infer $B \vee \neg B$). Since I endorse classical reasoning, I endorse relevant reasoning in the same sense—relevantly valid arguments also never take us from truth to untruth. But I think that relevantly valid arguments do more. I take it that, if you have a relevantly valid argument, and if the premises are true in a *situation*, then the conclusion is true in that situation, too. Relevantly valid arguments never take us from what is true-in-a-situation to something not true-in-that-situation. The inference from $A$ to $B \vee \neg B$ may because $B \vee \neg B$ may well fail in a situation (a situation may be incomplete) and explosion may fail because $A \wedge \neg A$ may be true in a situation (a situation may be inconsistent).[7]

Obviously, I need to make this much more precise if I am going to do work with the notion of truth-in-a-situation. But that would take us outside the scope of the present discussion. Suffice it to say that I think that the relevant notions can be clarified.[8] These gestures should be enough to explain why the two notions of logical consequence are of interest.

As an aside, let me indicate that the logical pluralism I endorse goes beyond accepting insights from classical and relevant logic. *Intuitionistic* logic can be understood as formalizing yet more of the pretheoretic notion of logical consequence: intuitionistically valid arguments are those such that a *construction* of the premises[9] is also a construction of the conclusion. Some arguments are classically (and relevantly) valid while being intuitionistically invalid. Discussing this as a *third* example would make this discussion needlessly longer and slightly more complicated. So I shall not consider intuitionistic logic further here, at least not in the body of the text.[10]

---

[5] So, I am not a *dialetheist* who thinks that some contradictions are true. For if I were, I would take it that some instances of explosion would take me from truth to untruth, at least if *some* claim were untrue.

[6] Cf. *Entailment. The Logic of Relevance and Necessity, Volumes 1 and 2* (Princeton: University Press, 1975 and 1992).

[7] None of this is intended to presume an answer, either positive or negative, to the issue of whether situations provide a *reduction* of relevant consequence. Nothing hangs on whether situations are parasitic on relevant consequence, or if relevant consequence is parasitic on situations, or if neither 'reduces' to the other.

[8] See John Barwise and John Perry, *Situations and Attitudes* (Cambridge: MIT, 1983); and my *op. cit.*

[9] This, too, is a pretheoretic notion formalized in a particular way: according to the Brouwer-Heyting-Kolmogorov account of construction. Cf. Michael Beeson, *Foundations of Constructive Mathematics. Metamathematical Studies* (Berlin: Springer, 1985); and Michael Dummett, *Elements of Intuitionism* (New York: Oxford, 1977).

## I

Rudolf Carnap is a natural hero to logical pluralists. For Carnap, logic is of fundamental importance to philosophical inquiry, but nothing about this inquiry (or about anything else) narrows the field of play to just one logic. For example, in *The Logical Syntax of Language*,[11] he wrote:

> The first attempts to cast the ship of logic off from the *terra firma* of the classical forms were certainly bold ones, considered from the historical point of view. But they were hampered by the striving after "correctness." Now, however, that impediment has been overcome, and before us lies the boundless ocean of unlimited possibilities (*ibid.*, p. xv).

This is stirring material. There is much for us to like in this paragraph. If we ought to investigate nonclassical accounts of logic, without concern for "correctness," then there is much scope for plurality. We may develop many different accounts of logical consequence, and let the dice of adequacy to our requirements in language building fall where they may. The realm of possibilities in logical theorizing is indeed boundless and unlimited. But *how* has the impediment of striving after "correctness" been overcome? For Carnap, the answer is the *principle of tolerance*.

> *It is not our business to set up prohibitions, but to arrive at conclusions* (*ibid.*, §17).

Carnap's aim is not to ban the choice of this or that logic, but to explore the consequences of such choices. This is expanded in the same section, where he writes:

---

[10] The complications arise from the fact that in intuitionistic logic some classical (and relevant) tautologies are no longer tautologies. For example, not every construction is a construction of $A \vee \neg A$. It is often thought that this means that intuitionists should reject some instances of the law of the excluded middle, or at the very least reject it as a general rule. This only follows if one is committed to rejecting nontautologies. But the pluralist need not have this commitment. After all, something might not be made true in every *construction* yet still be true in every possible world. Perhaps possible worlds are complete. I have written more on intuitionistic logic discussing these issues—cf. Beall and Restall, and my "Constructive Logic, Truth, and Warranted Assertibility," *Philosophical Quarterly*, LI (2001): 474-83.

[11] Amethe Smeaton, ed. (Paterson, NJ: Littlefield, Adams, 1959).

> *In logic, there are no morals.* Everyone is at liberty to build his own logic, i.e. his own form of language, as he wishes. All that is required of him is that, if he wishes to discuss it, he must state his methods clearly, and give syntactical rules instead of philosophical arguments (*ibid.*, §17).

For the Carnap of this period, all there is to logic is language adoption, and language adoption is radically unconstrained. If we wish to construct a language with rules like those of classical propositional logic, we may. If we wish to construct a language with weaker powers of inference, such as in a relevant logic, we may do that also. Carnap himself does this, by developing two systems, *Language I*, which embodies some constructive constraints, and *Language II*, which is a rich, classical type theory. This tolerance goes well beyond what we might feel comfortable to call *logic* today. For Carnap, languages could contain all sorts of inferential machinery, such as arithmetic and rich theories of types and functions—all the content of which was simply a part of the choice of the *language*, no matter how rich the mathematical theories were which have been adopted. The important constraint, if it is indeed a constraint, is that the language builder explain what principles are being used, in a clear, syntactic manner, as befits the choice of a language. If you wish to build a language containing a symbol, written '¬', that satisfies the rules of the classical account of negation, that is your prerogative. If, instead, you wish to build a language containing a symbol, also written '¬', that satisfies the rules of a relevant account of negation, that is fine, too. For Carnap, the choice of which language is to be preferred is pragmatic, to be answered by asking: What is the language *for*? Is your aim to give a comprehensive theory, able to prove as much as you can in a simple way? Classical negation might be for you. Is your aim to avoid contradiction and limit it if it arises? Then relevant negation might be for you.

Given this perspective of logic and language use, it is clear that the question '*Is this argument valid*?', when asked of a particular argument in a particular language, can be asked in two ways. It can be asked as an *internal question*, taking the choice of language of the argument for granted. In this case, it will be addressed by reference to the rules governing that language. Explosion is valid classically, but invalid relevantly. The question may also be asked *externally*. We can ask: "Should I choose a classical language, or a relevant one?" Here, the considerations do not give us an answer to the question 'Is this argument *really* valid?' but rather, they are reasons to adopt one language over another. After all, if we ask if a particular argument is valid or not, then if we know the language the argument is given in, the language itself will determine the validity or otherwise of the argument. For Carnap, to choose a logic just is to choose a language. So, we can see the question 'Ought I take this argument to be valid?' as asking something like this: 'Ought I *interpret* this argument as being in Language *X* or as in Language *Y*?'

It is unclear how far Carnap's tolerance may be defended in the light of Quinean criticisms of analyticity. Carnap is committed to a strong notion of analyticity, bound up with language construction and choice.[12] Valid inferences are correct *merely* in virtue of language adoption. Other truths expressible in the language do not become true in the same way. I shall not discuss how one might defend Carnap's views against these criticisms, for whether or not Carnap can be redeemed is irrelevant to the dialectic of the rest of the present discussion. In what follows, I shall need just a *part* of Carnap's doctrine, and this will be a part of Carnap's view that W. V. Quine happily adopts. It can put crudely as follows:

> Any difference in logical consequence is due to a difference in languages.

This is a commitment that Quine shares with Carnap, as we shall see, and it is enough to distinguish Carnap's pluralism from the kind of pluralism I wish to defend.[13]

It is clear, then, that Carnap's tolerance fosters a kind of pluralism about logical consequence which differs from the kind of pluralism I have presented in my introduction. For Carnap, diverging analyses of the validity of an argument result from diverging languages. The disputants in a dispute over the validity of an argument literally disagree in their *reading* of the argument. Given the disambiguation of the choice of a language, there can be no real disagreement, provided that the disputants know what language they speak. This is especially bad for a *pluralist*, who has appeared to have disagreed with herself. She is simply confused, equivocating from one language to another. Instead of saying that the argument is valid on one sense and invalid in another, she must be saying 'I can read the argument like *this* or I can read it like *that*'. I take it that this conclusion is unacceptable to my pluralism. If accepting different logics commits one to accepting different languages for those logics, then my pluralism is primarily one of *languages* (which come with their logics in tow) instead of *logics*. To put it graphically, as a pluralist, I wish to say that

---

[12] A fruitful discussion of Carnap's views can be found in Michael Friedman's *Reconsidering Logical Positivism* (New York: Cambridge, 1999). I am in broad agreement with Friedman's analysis of Carnap's views in *The Logical Syntax of Language*, though the detail of that analysis is not necessary for the argument of this paper.

[13] This is a commitment which is most likely very widely shared. John Etchemendy, for example, gives a strong statement of the view in his unpublished "Reflections on Consequence" (available from the author at http.//www-csli.stanford.edu/hp/Reflections.pdf) where he presents a view of logical consequence intimately tied to language identity. For Etchemendy, each language comes with one and only one consequence relation.

$$A, \neg A \vdash_C B \;\; \textit{but} \;\; A, \neg A \nvdash_R B$$

$A$ and $\neg A$ together, *classically* entail $B$, but $A$ and $\neg A$ together do not *relevantly* entail $B$. On the other hand, Carnap wishes to say that

$$A, \neg_C A \vdash B \;\; \textit{but} \;\; A, \neg_R A \nvdash B$$

$A$ together with its *classical* negation entails $B$, but $A$ together with its *relevant* negation need not entail $B$.[14]

The domain of variation differs considerably between analyses.[15] For Carnap, it is not the one argument form which we evaluate differently—it is two different argument forms. In what follows, I shall ask whether Carnap's analysis is mandatory, or whether there is scope for the pluralist to admit different consequence relations over the one language.

## II

There might be a number of reasons why you might take the proponents in a disagreement over logical consequence (such as that over explosion or disjunctive syllogism) to use different languages. In this section, I shall examine *three* reasons, and show that none of them leads to insurmountable problems for the pluralist. In doing so, we shall see what resources the thoroughgoing pluralist can have for admitting more than one consequence relation on the one language. The first fully-developed argument to the effect that a difference in logical principles marks a difference in language is to be found in Quine's work. The most well-known presentation of the argument is in his *Philosophy of Logic*.[16]

---

[14] The same story could be told in the case of intuitionistic logic. Instead of taking the intuitionistic *language* to differ from a classical language, we can identify them. The difference between Heyting arithmetic (which takes the axioms of classical Peano arithmetic, and considers their consequences in intuitionistic predicate logic) and classical Peano arithmetic is then what it appears to be on the surface. Some theorems of Peano arithmetic are not theorems of Heyting arithmetic, but all theorems of Heyting arithmetic are also theorems of Peano arithmetic. If we were to take Carnap's doctrine literally, this could not be the case. For Carnap, all theorems of Heyting arithmetic featuring 'intuitionistic negation' would not be expressible in the language of classical arithmetic, which features no such connective. This interpretation is at the very least, *strained*. It is common in metamathematical practice to consider the relative strengths of theories couched in different *logics*.

[15] Carnap *at least* wants to say this. He probably also wishes to say that in the first consecution, the $A$ and $B$ range over sentences in language $C$ and in the second they range over sentences in language $R$. As Otávio Bueno has pointed out in conversation, however, the later Carnap—in "Observation Language and Theoretical Language," in *Rudolf Carnap, Logical Empiricist: Materials and Perspectives*, Jaakko Hintikka, ed. (Boston: Reidel, 1975), pp. 75-85—might not be committed to this view.

> To turn to a popular extravaganza, what if someone were to reject the law of non-contradiction and so accept an occasional sentence and its negation as both true? An answer one hears is that this would vitiate all science. Any conjunction of the form 'p $\wedge \neg$p' logically implies every sentence whatever; therefore acceptance of one sentence and its negation as true would commit us to accepting every sentence as true, and thus forfeiting all distinction between true and false.
>
> In an answer to this answer, one hears that such a full-width trivialization could perhaps be staved off by making compensatory adjustments to block this indiscriminate deducibility of all sentences from an inconsistency. Perhaps, it is suggested, we can so rig our new logic that it will isolate its contradictions and contain them.
>
> My view of this dialogue is that neither party knows what he is talking about. They think they are talking about negation, '$\neg$', 'not'; but surely the notation ceased to be recognizable as negation when they took to regarding some conjunctions of the form 'p $\wedge \neg$p' as true, and stopped regarding such sentences as implying all others. Here, evidently, is the deviant logician's predicament: when he tries to deny the doctrine he only changes the subject (*ibid.*, p. 81).

The particular view presented here goes back many years. Hints are first found in "Truth by Convention" (written in 1936[17]) and "Carnap and Logical Truth" (written in 1954[18]). In "Truth by Convention," he considers the consequences of adding a one place connective '$\neg$' to a language (together with other logical machinery) in order to show that a mere convention suffices to account for the truth of the theorems of logic.[19] In this discussion, a "mark in the margin" indicates that the corresponding expression is taken to be true.

> If we make a mark in the margin opposite an expression '—', and another opposite '$\neg$—', we sin only against the established usage of '$\neg$'

---

[16] Englewood Cliffs, NJ: Prentice-Hall, 1970

[17] Reprinted in *The Ways of Paradox and Other Essays* (New York: Random House, 1966), pp. 70-99.

[18] Reprinted in *The Ways of Paradox and Other Essays*, pp. 100-25.

[19] He then shows that you can do this for *any* term, making any truth a truth by convention, evacuating any distinction you can draw by calling something "true by convention."

> as a denial sign. Under the latter usage '—' and '¬—' are not both true; in taking the both by convention as true we merely endow the sign '¬', roughly speaking, with a meaning other than denial.[20]

So, violating the principle of noncontradiction by taking an expression and its '¬- negation' to both be true simply rules out the interpretation of '¬' as negation. In "Carnap and Logical Truth," the discussion shifts to considering the issue of a disagreement between two people ostensibly sharing a language.

> When someone disagrees with us as to the truth of a sentence, it often happens that we can convince him by getting the sentence from other sentences, which he does accept, by a series of steps each of which he accepts. Disagreement that cannot be thus resolved I shall call *deductively irresoluble*. Now if we try to warp the linguistic doctrine of logical truth around into something like an experimental thesis, perhaps a first approximation will run thus: *Deductively irresoluble disagreement as to a logical truth is evidence of deviation in usage (or meanings) of words.*[21]

These are early graspings toward an argument which finds its maturity in *Word and Object*.[22] Here, Quine formulates the scenario of *radical translation*, where he considers what evidence is to hand when a field linguist seeks to translate a completely unknown language into English. His claim is that the available evidence (the correlation of utterances and stimulus situations) does not pin down a single translation manual as favored. Of course, there are certain canons of translation which ought to be followed, if we are to have an "acceptable" translation, but these canons do not fix a single translation, and so, they do not fix "meaning." The major principle is the principle of charity. When translating, the field linguist presumes that the language users are rational and are not generally mistaken about everything.

> The semantic criterion of negation is that it turns any short sentence to which one will assent into a sentence from which one will dissent, and vice versa....
>
> The maxim of translation underlying all this is that assertions startlingly false on the face of them are likely to turn on hidden differences of language. This maxim is strong enough in all of us to swerve us even from the homophonic method that is so fundamental to the very acquisition and use of one's mother tongue.

---

[20] "Truth by Convention," p. 90

[21] "Carnap and Logical Truth," p. 105

[22] Cambridge: MIT, 1960.

> The common sense behind the maxim is that one's interlocutor's silliness, beyond a certain point, is less likely than bad translation—or, in the domestic case, linguistic divergence. Another account of the matter, as it touches logical laws in the domestic case, is as follows. The logical particles 'and', 'all', etc. are learned only from sentential contexts. Dropping a logical law means a devastatingly widespread unfixing of truth values of contexts of the particles concerned, leaving no fixity to rely on in using those particles (*ibid.*, pp. 57, 59, 60).

As this quote shows, Quine applies these results in generality. They are not restricted to the situation of the field linguist, since no one, when learning a language, has access to any evidence that the field linguist lacks.

The presentation is somewhat more muted in Quine's recent book, *Pursuit of Truth*,[23] but the general point remains.

> Our linguist then goes on tentatively identifying and translating observation sentences. Some of them are perhaps compounded out of others of them, in ways hinting of our logical particles 'and', 'or', 'but', 'not'. By collating the situations that command the natives' assent to the compounds with the situations that command assent to the components, and similarly for dissent, the linguist gets a plausible line on such connectives (*ibid.*, pp. 44, 45).

In the rest of this section, I shall assume that Quine's general thesis is sound; that the available evidence underdetermines appropriate translation. In this way, I shall not attempt to solve the problem by being anti-Quinean. The task then, is to understand what is going on in a disagreement over disjunctive syllogism, or over explosion. Does any such disagreement involve a bifurcation of languages?

Consider the case where we have a classical reasoner (for simplicity, we shall call our chosen reasoner "Quine") who tries to interpret my utterances, as a pluralist who *at least* adopts relevant consequence. Now, Quine can question me as much as he likes, asking whether I agree or disagree with various sentences. And generally speaking, if I agree with '*p*', I shall disagree with 'not *p*', and vice versa. Even if he tries asking me tricky paradoxical sentences, like his favorite version of the liar paradox:

> 'Yields falsehood appended to its quotation' yields falsehood appended to its quotation.

I shall simply demur, as I do not have a settled opinion on the truth values of paradoxical sentences like these. Quine will eventually have to agree that my use of 'not' agrees with his, insofar as the manipulation of the values of assenting and dissenting goes, as it meets all of the requirements he gives for a connective to be a negation.[24]

---

[23] Cambridge: Harvard, 1992, second edition.

We are not home yet, however. Recall that Quine said in *Philosophy of Logic* that it was as soon as the interlocutors "stopped regarding such sentences [contradictions] as implying all others" that the discussion stopped being about negation. Well, if Quine wants to ask me, he will see that I do not think that contradictions generally imply *everything*. (Interestingly, this consideration appears only in *Philosophy of Logic*, not in his works where he develops the translation argument in some detail.) Does this mean that my use of negation diverges from his? I think it need not. One reason for this is Quine's own use of terms like 'imply'. For Quine, the only conditional worth talking about (or at least, theorizing about) is the material conditional. That is, a conditional of the form 'if $p$ then $q$', if it is to be given any truth conditions at all, is to be read as simply saying 'either not $p$ or $q$'. For Quine, implication is a derivative notion, and it plays no important place in his theorizing.[25] So what should Quine deduce when I deny some of his claims of the form 'if $p$ and not $p$ then $q$'? He should probably conclude that my usage of 'if...then' differs from his, but this would not be surprising. This impression would be strengthened if he went on to ask me the question, phrasing the conditional in its material form. For to any claim of the form 'not ($p$ and not $p$) or $q$' I shall agree, simply because I always agree that '$p$ and not $p$' is false. So Quine ought to conclude that my use of the truth functions is completely "classical." We do not disagree in our reading of the negation connective in an argument like explosion or disjunctive syllogism. The *Quinean* route to language variance in the fragment of our language including at least the Boolean connectives of conjunction, disjunction, and negation is blocked. But it was just this language in which different consequent relations reside. So a Quinean argument is not sufficient to close the door to pluralism.[26]

---

[24] Of course, even some monists who accept classical logic think that "not" and denial come apart (cf. Jamie Tappenden, "Negation, Denial and Language Change in Philosophical Logic," in Gabbay and Wansing, eds , pp. 261-98), so if we wish to reject Quine's close connection between negation and denial, we can do so in a way that does not beg the question against the classical monist. I am happy to keep the correlation between negation and denial, and so I can expect that Quine would find my 'not' used to express negation.

[25] As any cursory checking of the indices of his books for words like 'implication', 'consequence', 'entailment', or 'conditional' will attest

[26] The argument for pluralism with respect to classical and intuitionistic logic is similar, yet it has distinctive features of its own The kind of pluralism supported by

Quine's considerations about translation are not the only motivations for thinking that differences in logical consequence give rise to differences in languages. Another charge against the possibility of pluralism comes from the quarter of truth-conditional semantics. If 'not' in the mouth of the classical reasoner has different truth conditions to 'not' in the mouth of the relevant reasoner, then if the meaning of a particle like 'not' consists in its truth conditions, then when a classical reasoner and a relevant reasoner both say 'not' they mean different things—and thereby speak different languages.

Do the proponent and opponent of a disagreement over disjunctive syllogism or explosion agree about the truth conditions of statements of the form of $\neg A$? It seems obvious that they do not. After all, the classical reasoner will happily assent to claims in the semantics of a logic of the form:

$\neg A$ is true in $x$ if and only if $A$ is not true in $x$

And the relevant reasoner, famously, will prefer other claims analyzing the behavior of negation. There are many different candidate analyses. The one I prefer[27] goes like this:

$\neg A$ is true in $x$ if and only if $A$ is not true in $y$ for any $y$ compatible with $x$

These are different analyses of negation. Does it not follow that different connectives are being analyzed?

This need not follow at all. Granting, for the moment, that *meaning* identity requires identity of *truth conditions*, it is not obvious that the truth conditions of negation must be analyzed differently by the relevant reasoner and the classical reasoner. There are a number of options available for the pluralist. I shall examine just two.

*Minimalist truth conditions*: first, we could be *minimalist* about truth conditions. We could agree with Donald Davidson[28] and others that

---

this argument will not be congenial to the intuitionist who rejects some instances of the truth-functional tautology $A \vee \neg A$. As indicated in footnote 10, however, not all adherents of intuitionistic logic need reject the law of the excluded middle. It may be agreed that it is not constructed by all constructions, and hence it is not a tautology (of intuitionistic logic); but it is a further step to conclude that it is not *true*, or to reject it. The analogy holds with relevant consequence and *dialetheism*. We may agree that contradictions do not entail everything without thereby being committed to occasionally *accepting* contradictions.

[27] Cf. Dunn, "Star and Perp: Two Treatments of Negation," in James E. Tomberlin, ed., *Philosophical Perspectives, Volume 7* (Atascadero, CA: Ridgeview, 1994), pp. 331-57; and my "Negation in Relevant Logics."

[28] See, for example, his "In Defense of Convention T," in *Inquiries into Truth and Interpretation* (New York: Oxford, 1984), pp. 65-75.

truth conditions are bare biconditionals connecting sentences and the predications of truth, such as this:

$\neg A$ is true if and only if $A$ is not true

The classical and relevant reasoners can agree that this biconditional is true, and if this is all there is to say about the meaning of $\neg$, we have no disagreement over its meaning. More must be said of how the two clauses presented by our partners relate to the meaning constraint given by the T-biconditional. One plausible way of doing this is to take the clause '$\neg A$ is true in $x$ if and only if $A$ is not true in $x$' as quantifying over "worldlike" objects while the "relevant" clause '$\neg A$ is true in $x$ if and only if $A$ is not true in $y$ for any $y$ compatible with $x$' quantifies over "situationlike" objects. Given that $\neg$ has the natural truth conditions, we can take the semantic clauses in formal model theories of different logics as picking out different kinds of objects.

*Maximalist truth conditions*: perhaps we need more than simple bare T-biconditionals as truth conditions. Perhaps we need truth conditions to say more about the different "worlds" or "situations" in which claims are true. Another way we can find harmony between the classicalist and the relevantist is to take the truth conditions of negation to be the collection of *all* of the semantic clauses used, in either the classical or the relevant semantics. All we need to do to ensure that we have no *difference* of meaning is to ensure that the two accounts do not clash; but they do not. If a "world" is simply a consistent and complete "situation," then the two accounts agree where they overlap.[29] The most disagreement that we have is that the classical logician does not use as *comprehensive* a part of the truth conditions of negation as does the relevantist. To show that classical and relevant negation are connectives with different truth conditions, the monist must now show that one of the accounts is defective, and this is no longer an argument against pluralism per se but an argument against either the classical or the relevant account of negation. As such, I can leave it aside.

It follows that we have more than enough ways to reconcile the claim that meaning identity requires identity of truth conditions with pluralist accounts of logical consequence. Concern for truth conditions gives us no reason to reject pluralism.

I shall end this section by examining one more argument against the claim that the pluralist can be univocal. Do two different logical consequence relations agree about the *inferential role* of statements of the form $\neg A$? It seems obvious that they do not. I shall attempt to argue that this is less obvious than it might first appear.

---

[29] Cf. my "Negation in Relevant Logics."

I shall not adjudicate between non-normative[30] and normative[31] inferential-role accounts of the meanings of logical vocabulary, such as negation, as the point of this argument can be made equally well according to either account, and so can the pluralist's response to the objection. First consider Robert Brandom's inferentialism (*ibid.*). For Brandom, questions of meaning are questions of commitment and entitlement, and these are essentially *normative* questions. We can explicate the inferential role of negation as follows: commitment to $\neg A$ is incompatible with entitlement to $A$, and entitlement to $A$ is incompatible with commitment to $\neg A$. Note that nothing here has decided the issue between proponents of classical or relevant consequence. The classicalist and the relevantist can agree about how commitment and entitlement relate to negation.

Given that basic agreement, different logics may well relate to entitlement in different ways. For example, it is plausible that you are never entitled to the premises and the negation of the conclusion of a classically valid argument. For example, you are never entitled to $(A \wedge \neg A) \wedge \neg B$, simply because one is never entitled to the first conjunct $A \wedge \neg A$. A relevant logic is a much more plausible candidate for the "logic" of commitment preservation. Given a relevantly valid argument, if you are committed to the premises, you are, in virtue of this, committed to the conclusion, too. In this way we can agree that no one is really *entitled* to a contradiction $A \wedge \neg A$ (classical logic as a logic of entitlement rules that out) while still maintaining that commitment to a contradiction does not bring along with it commitment to anything and everything. It is quite natural for an inferentialist like Brandom to be a pluralist about logical consequence.[32] Much more needs to be said to make this a clear presentation of the notions of logical consequence arising out of Brandom's analysis of meaning, but the basic point has been made. He is not an *inferentialist* in the sense of someone using a basic relation of inference to determine the meanings of parts of our language. No, Brandom uses the basic notions of commitment and entitlement to do this job. None of this picks out a unique fundamental relation of logical consequence. An inferentialist such as Brandom may be a pluralist about logical consequence.

---

[30] Cf. Gilbert Harman, *Change in View: Principles of Reasoning* (Cambridge: MIT, 1986).

[31] Cf. Robert Brandom, *Making It Explicit* (Cambridge: Harvard, 1994).

[32] For details of how to make explicit the "logics" of commitment and entitlement, on Brandom's account, the papers by Mark Lance and Philip Kremer are good places to start—Lance, "Two Concepts of Entailment," *Journal of Philosophical Research*, xx (1995): 113-37; Lance and Kremer, "The Logical Structure of Linguistic Commitment II: Systems of Relevant Commitment Entailment," *Journal of Philosophical Logic*, xxv (2000): 425-49.

To end this section, I shall consider a more difficult case for the pluralist: Gilbert Harman's account of the meanings of logical constants (*op. cit.*). For Harman, the meaning of the logical constants is to be found in the *basic* valid inferences in which they feature. In this way, despite the fact that $\neg(\neg A \vee \neg B)$ and $A \wedge B$ are inferentially equivalent, they have different meanings, because the *basic* inferences involving $\neg(\neg A \vee \neg B)$ differ from those involving $A \wedge B$, as they must use negation and disjunction rules, not conjunction rules. If we are a pluralist about consequence, what can we make of Harman's analysis? It appears that a relevantist and a classicalist do not share the one class of basic inferences involving negation, and so, a meaning change threatens. Again, there seem to be a number of possible responses the pluralist can make.

*"Inferences" and logical consequence*: one response is that the kinds of *inference* required to fix meaning is not logical consequence at all, but something else, such as dispositions to believe, or perhaps something normative. If we read "inference" in this account as not picking out a kind of logical consequence, then the fact that inference plays such a central role in determining meaning is not a problem for the pluralist. This is the adaptation for Harman of our response to criticism based on Brandom's inferentialism. The basic "inferences" are not themselves a logic, and logical consequence relations can be built out of these basic "inferences" in a number of different ways.

While this is a satisfying response to Brandom's inferentialism, it is strained in this context. For Harman, the meaning determining inferences featuring negation just *are* basic natural deduction inferences from a proof theory of a particular logic.[33]

*Pick one consequence relation*: if we grant, with Harman, that each connective has a meaning, and such a meaning is to be found with its inferential role analyzed as the basic inferential moves in a natural deduction system, the other response available is to bite the bullet. We may admit that the basic rules do pick out one consequence relation. Again, as with the case of truth conditions, there is a *mini-*

---

[33] Harman, "The Meanings of Logical Constants," in Ernest Le Pore, ed., *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson* (New York: Blackwell, 1986), pp. 125-34.

*malist* and a *maximalist* response.[34] The maximalist response takes the meaning determining rules to be those of classical logic, the minimalist response takes meaning to be determined by the weaker rules of relevant consequence. The point is the same with both responses. We can pick *one* of our plurality of consequence relations as determining the meaning of a connective. If we take the primitive classical negation rules as the ones used to determine the meaning of negation, then we take the weaker relevant rules to tell us something else about the behavior of *that* connective. If any set of rules is sufficient to pick out a single meaning for the connective, take that set of rules, and accept those as meaning determining. The other rules are important when it comes to giving an account of a kind of logical consequence, but they are not used to determine meaning.

Does this threaten pluralism about logical consequence? I agree that the response is *odd*. It does seem to single out a single consequence relation as best for a particular job—the job of elucidating meanings.[35] The fact that a particular consequence relation is best suited for some job or other does not mean that we have become less pluralist. The doctrine is that each logical consequence relation is equally *logical*, not that each consequence relation is equally good for everything. I already agree that one logic (a logic which is classical in the language of conjunction, disjunction, and negation) is best for necessary truth preservation; that another is best for considerations of relevance. If one is best for cashing out the meanings of sentence operators, then so be it.

*Inference schemes*: there is one more radical approach to the problem. The crucial idea here is that the basic properties of negation are indeed encoded in inference patterns like these:

$$\frac{X \vdash A, Y}{X, \neg A \vdash Y} \qquad \frac{X, A \vdash Y}{X \vdash \neg A, Y}$$

Here, the inference patterns determines the meaning of the connective.[36] The meaning is presented in the two rules. The *variation* in the rules appears in the behavior of the *structural rules* applied in proofs.[37] In classical logic, the comma contains the full complement of structural rules, and both $X$ and $Y$ are allowed to be nonempty. In intuitionistic logic, all structural rules are allowed, but we do not allow the consequent of a consecution to contain more than one formula. In relevant logic, we do not allow weakening: that is, we do not allow the inference from $X \vdash Y$ to $X, A \vdash Y$ or to $X \vdash Y, A$, on grounds of *relevance*. Consider the following proof of the relevantly invalid explosion $A, \neg A \vdash B$:

---

[34] But here the *maximalist* response is broadly classical and the *minimalist* response is more nonclassical, in contrast to the case with truth conditions.

[35] Actually, this does not quite follow. It simply picks out a single set of inference rules for each part of the language that is given an inferential account of meaning. There is nothing to say yet that each set of inference rules must come from the same logic. It is hard to see, however, how or why the rules might come from different consequence relations.

[36] In these proofs, $X \vdash Y$ is read as saying that, if everything in $X$ is true, something in $Y$ is true. So if everything in $X$ and $A$ is true, then something in $Y$ is true, we can infer that, if everything in $X$ is true, then either something in $Y$ is true, or $\neg A$ is true.

$$\dfrac{\dfrac{A \vdash A}{A \vdash A, B}}{A, \neg A \vdash B}$$

The relevant logician complains about the *first* step, not the second. A relevantist is happy to infer $A, \neg A \vdash B$ from $A \vdash A, B$, but is *not* happy to infer $A \vdash A, B$ from $B$.[38] The different logics considered here differ in structural rules, not in our theory of negation.

So plurality is allowed in *applications* of inference schemes, but the schemes determining the meaning of connectives are unitary. We have another response to the argument that inferential theories of meaning lead us to monism about logical consequence in a given language.[39]

Whatever response we take at this point (and I think that the response in terms of structural rules is the most inviting) we have seen that we can be pluralist while accepting that meaning is conveyed by inferential role.

So three major considerations about meaning and language identity have been shown insufficient to warrant Carnap's conclusion.

---

[37] Cf. Dunn, "A 'Gentzen' System for Positive Relevant Implication," *Journal of Symbolic Logic*, xxxviii (1974): 356-57; Belnap, "Display Logic," *Journal of Philosophical Logic*, xi (1982): 375-417; John K. Slaney, "A General Logic," *Australasian Journal of Philosophy*, lxviii (1990): 74-88; my *An Introduction to Substructural Logics* (New York: Routledge, 2000).

[38] Of course, this is because of the difference in the way the relevantist and the classicalist treat the form of premise and conclusion combination, here both expressed by the comma.

[39] This can be seen as a version of either of the two responses we have already seen. If you take a scheme to not be an inference, then we agree that the meaning determining rules are not themselves inferences. If, instead, you take the scheme to be its *weakest* instance with the fewest structural rules, then the scheme is a particular set of basic inferences, which may be expanded on by the means of other structural rules.

Choosing a language is freer than Carnap would have us think. Given a language, considerations of meaning are not sufficient to narrow down the number of genuine consequence relations on that language down to one.

Carnap's principle of tolerance is inviting for the pluralist, but it is ultimately unsatisfying. Different accounts of logical consequence need not be restricted to relationships *between* languages. They can also occur *within* languages.

GREG RESTALL

University of Melbourne

## NOTES AND NEWS

The Editors regret the death of Robert Nozick on January 23, 2002, at the age of 63. Nozick obtained an A.B. degree from Columbia College in 1959, and M.A. and Ph.D. degrees from Princeton University in 1961 and 1963, respectively. After teaching at Princeton and Rockefeller University, Nozick came to Harvard University as a full professor in 1969. He became Arthur Kingsley Porter Professor of Philosophy in 1985, and in 1998 was named the Joseph Pellegrino University Professor. His published works include *Anarchy, State, and Utopia* (1974), *Philosophical Explanations* (1981), *The Examined Life* (1989), *The Nature of Rationality* (1995), *Socratic Puzzles* (1997), and *Invariances: The Structure of the Objective World* (2001).

Professor Nozick was the recipient of many awards and honors, among them the Presidential Citation from the American Psychological Association in 1998. He was also a fellow of the American Academy of Arts and Sciences, a member of the Council of Scholars of the Library of Congress, a corresponding fellow of the British Academy, and a senior fellow of the Society of Fellows at Harvard. He served as the president of the American Philosophical Association's Eastern Division (1997-1998), was a Christensen visiting fellow at St. Catherine's College, Oxford University (1997), as well as a cultural adviser to the U.S. Delegation to the UNESCO Conference on World Cultural Policy (1982), and he delivered the six John Locke Lectures at Oxford University (1997). He held fellowships from the Guggenheim Foundation, the Rockefeller Foundation, the National Endowment for the Humanities, and the Center for Advanced Study in the Behavioral Sciences at Stanford University.

Nozick's controversial and challenging views gained him considerable attention and influence in the world beyond the academy. A former member of the radical left who was converted to a libertarian perspective as a graduate student, he contributed to his discipline's understanding of individual rights in a way that lionized the individual rather than the collective.

# Language and Philosophy

NEW FROM
CSLI
PUBLICATIONS

Founded in 1983, the Center for the Study of Language and Information launched its publishing program under the imprint of CSLI Publications to promote scholarship in linguistics, logic, philosophy, psychology, computer science, and other fields in cognitive science that contribute to the development of theories of language and information. The University of Chicago Press distributes nearly all CSLI Publications titles.

## INTERPRETING DAVIDSON

**Edited by Petr Kotatko, Peter Pagin, and Gabriel Segal**

Among the most important philosophers of mind and language in recent decades, Donald Davidson articulated "anomalous monism" as well as ideas for unifying the general theory of linguistic meaning with semantics for natural language. This book collects original essays on Davidson by his leading contemporaries; Davidson replies to each and presents an original paper.

PAPER $27.50

## EXPLAINING BELIEFS

Lynne Rudder Baker and Her Critics

**Edited by Anthonie Meijers**

The philosophy of mind has long been dominated by the view that mental states are identical with, constituted by, or grounded in brain states. Lynne Rudder Baker has been a persistent critic of this view, developing instead a theory grounded in Practical Realism. This first critical book-length evaluation of her views and criticism allows leading philosophers to answer her challenge and explore the consequences of Practical Realism; Baker herself provides thoughtful replies.

PAPER $27.50

## THE USE OF LANGUAGE

**Prashant Parikh**

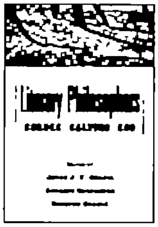
Building on the work of J. L. Austin and Paul Grice, *The Use of Language* develops an original and systematic game-theoretic account of communication, speaker meaning, and addressee interpretation, extending this analysis to conversational implicature and the Gricean maxims, illocutionary force, miscommunication, visual representation and visual implicature, and aspects of discourse.

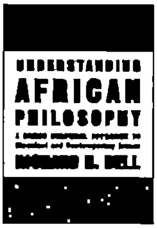
PAPER $24.50

DISTRIBUTED BY THE UNIVERSITY OF CHICAGO PRESS
1427 East 60th Street, Chicago, IL 60637 • www.press.uchicago.edu

ROUTLEDGE

NEW IN PHILOSOPHY





NEW IN THE *THINKING IN ACTION* SERIES

*Series editors, Simon Critchley and Richard Kearney*

**ON HUMOUR**
*Simon Critchley*

**ON LITERATURE**
*J. Hillis Miller*

**ON STORIES**
*Richard Kearney*

**ON FILM**
*Stephen Mulhall*

**UNDERSTANDING AFRICAN PHILOSOPHY**
A Cross-cultural Approach to Classical and Contemporary Issues
*Richard H. Bell*

*Understanding African Philosophy* serves as a critical guide to some of the most important issues in modern African philosophy. Bell introduces readers to the complexity of Africa, the legacy of colonialism, the challenges of post independence Africa, and other recent developments in African philosophy.

208 pp

**LITERARY PHILOSOPHERS**
Borges, Calvino, Eco
*Edited by Jorge J.E. Gracia, Carolyn Korsmeyer and Rodolphe Gasché*

*Literary Philosophers* brings together a group of distinguished philosophers, literary scholars, and comparativists to explore and debate the relationship between philosophy and literature in the works of Borges, Calvino, and Eco.

328 pp

**NEW BRITISH PHILOSOPHY**
The Interviews
*Edited by Julian Baggini and Jeremy Stangroom*

Clear and engaging, *New British Philosophy* contains sixteen fascinating interviews with some of the top philosophers working in Britain today, on topics that range from music to the mind and feminism to the future of philosophy.

320 pp

**HISTORY OF IRISH THOUGHT**
*Thomas Duddy*

*"Strikingly original and sorely needed. There is an engaged, illuminating intelligence at work here, reflected in the crisp, remarkably lucid style."*

—Terry Eagleton, author of *The Truth About the Irish*

304 pp

www.routledge-ny.com

1.800.634.7064

ROUTLEDGE

## Outstanding Scholarship from Cambridge

### On the Trinity
*Augustine*
*Gareth Matthews, Editor*
Cambridge Texts in the History of Philosophy
0-521-79231-2 Hardback $55 00
0-521-79665-2 Paperback $20 00

### Computability and Logic
Fourth Edition
*George S. Boolos, John P. Burgess, and Richard Jeffrey*
0-521-80975-4 Hardback $75 00
0-521-00758-5 Paperback $28 00

### Deontic Morality and Control
*Ishtiyaque Haji*
Cambridge Studies in Philosophy
0-521-81387-5 Hardback $65 00

### Herder: *Philosophical Writings*
*Johann Gottfried von Herder*
*Michael Forster, Editor*
Cambridge Texts in the History of Philosophy
0-521-79088-3 Hardback $60 00
0-521-79409-9 Paperback $23.00

### Interpreting Probability
Controversies and Developments in the Early Twentieth Century
*David Howie*
Cambridge Studies in Probability, Induction and Decision Theory
0-521-81251-8 Hardback $60.00

### Kant's Search for the Supreme Principle of Morality
*Samuel J. Kerstein*
0-521-81089-2 Hardback $60 00

### The Correspondence Theory of Truth
An Essay on the Metaphysics of Predication
*Andrew Newman*
Cambridge Studies in Philosophy
0-521-81139-2 Hardback $60.00

### The Cambridge Translations of Medieval Philosophical Texts
Volume 3: Mind and Knowledge
*Robert Pasnau, Editor*
The Cambridge Translations of Medieval Philosophical Texts
0-521-79356-4 Hardback $95 00
0-521-79795-0 Paperback $35.00

### Philosophy and German Literature, 1700-1990
*Nicholas Saul, Editor*
Cambridge Studies in German
0-521-66052-1 Hardback $65.00

### The Culture of Morality
Social Development, Context and Conflict
*Elliot Turiel*
0-521-80833-2 Hardback $32.00

### Religion and the Obligations of Citizenship
*Paul J. Weithman*
0-521-80857-X Hardback $55 00

MORALITY

*Available in bookstores or from*
CAMBRIDGE
UNIVERSITY PRESS
800-872-7423
us.cambridge.org/philosophy

# Classic Works in the History of Ideas

Thoemmes Press's high-quality, multi-volume editions bring together the major writings of important figures, works that exemplify key themes and debates, and the lesser-known and neglected works of the past—including modern classics previously out of print and indispensable reference works that are publishing firsts.

For more information about these and other Thoemmes Press books, visit the Thoemmes Press web site at www.thoemmes.com.

In North America, Thoemmes Press books are distributed by the University of Chicago Press.

**ESSAYS AND TREATISES ON SEVERAL SUBJECTS**
**David Hume**
1130 PAGES, 2 VOLUMES 1-85506-935-0
CLOTH $265.00

**THE JOURNAL OF SPECULATIVE PHILOSOPHY, 1867–1893**
9240 PAGES, 22 VOLUMES 1-85506-960-1
CLOTH $2150.00

**KANT BIOGRAPHIEN**
**Edited by Manfred Kuehn**
2335 PAGES, 8 VOLUMES 1-85506-959-8
CLOTH $895.00

DISTRIBUTED BY
THE UNIVERSITY OF CHICAGO PRESS
1427 East 60th Street, Chicago, Illinois 60637 • www.press.uchicago.edu

New from Thoemmes Press

**BIOGRAPHIA PHILOSOPHICA**
Being an Account of the Lives, Writings, and Inventions of the Most Eminent Philosophers and Mathematicians
**Benjamin Martin**
567 pages 1-85506-973-3
Cloth $140.00

**PHILOSOPHICAL AND THEOLOGICAL WRITINGS**
**John Norris**
Edited by Richard Acworth
2750 pages, 8 Volumes 1-85506-901-6
Cloth $950.00

**THE IRISH ENLIGHTENMENT AND COUNTER-ENLIGHTENMENT**
**Edited by Patricia O'Riordan and David Berman**
2490 pages, 6 Volumes 1-85506-959-3
Cloth $590.00

**JOSIAH ROYCE'S LATE WRITINGS**
A Collection of Unpublished and Scattered Works
**Josiah Royce**
Edited by Frank M. Oppenheim
620 pages, 2 Volumes 1-85506-911-3
Cloth $260.00

**HUME'S RECEPTION IN EARLY AMERICA**
**Edited by Mark G. Spencer**
440 pages, 2 Volumes 1-85506-934-2
Cloth $265.00

**AN ESSAY TOWARDS A REAL CHARACTER AND A PHILOSOPHICAL LANGUAGE**
**John Wilkins**
638 pages 1-85506-941-5
Cloth $295.00

**STATEMENT AND INFERENCE**
**John Cook Wilson**
1091 pages, 2 Volumes 1-85506-958-X
Cloth $260.00

**FOUNDATIONS IN THE HISTORY OF PSYCHOLOGY**
**Edited by Robert H. Wozniak**
2500 pages, 7 Volumes 1-85506-940-7
Cloth $745.00

DISTRIBUTED BY
THE UNIVERSITY OF CHICAGO PRESS
1427 East 60th Street, Chicago, Illinois 60637 • www.press.uchicago.edu

**FUNCTIONS**
*New Essays in the Philosophy of Psychology and Biology*
Edited by ANDRE ARIEW, ROBERT CUMMINS, and MARK PERLMAN

In explaining aspects of the natural world, including the aspects of mind, scientists have frequently used the concept of function. But what are functions? Here, 15 leading scholars of philosophy of psychology and philosophy of biology present new essays on functions.

June 2002 450 pp , 6 illus paper $24 95 cloth $70 00

*New in paperback!*
**QUESTIONS OF TIME AND TENSE**
Edited by ROBIN LE POIDEVIN

"A remarkable collection of essays ....[a] well-produced volume."
—*Choice*.

This book brings together new essays on a major focus of debate in contemporary metaphysics: does time really pass, or is our experience of time as consisting of past, present, and future an illusion?

1998 (paper 2002) 312 pp , paper $22 95 cloth $65 00

**SUFFERING AND MORAL RESPONSIBILITY**
JAMIE MAYERFELD

"provocative and insightful...."
—*Choice*.

Understanding suffering in hedonistic terms as an affliction of feeling, Mayerfeld addresses difficulties associated with its identification and measurement. He then turns to an examination of the duty to relieve suffering. its content, its weight relative to other moral considerations, and the role it should play in our lives.

1999 (paper 2002) 240 pp paper $21 75 cloth $49 95

**THE MYTH OF OWNERSHIP**
*Taxes and Justice*
LIAM MURPHY and THOMAS NAGEL

Taxes arouse strong passions, fueled not only by conflicts of economic self-interest, but by conflicting ideas of fairness. Taking up ethical issues about individual liberty, interpersonal obligation, and both collective and personal responsibility, Murphy and Nagel force us to reconsider how our tax policy shapes our system of property rights.

2002 240 pp $25 00

**ENVIRONMENTAL JUSTICE**
*Creating Equity, Reclaiming Democracy*
KRISTIN SHRADER-FRECHETTE

Indigenous people, minorities, and the poor are more likely to live near toxic wastes, breathe dirty air, and work in hazardous jobs. Shrader-Frechette argues for these individuals' right to environmental justice: to equalization of the burdens of pollution, noxious development, and resource depletion.

August 2002 304 pp $35 00

OXFORD

**THEORY AND TRUTH**
*Philosophical Critique within Foundational Science*
LAWRENCE SKLAR
"Exemplary, and always anchored in the concrete details of real scientific theorizing....[It] demonstrate[s] the virtues of a sober and careful appraisal of the issues concerning realism about scientific theories.."—Peter Menzies, *TLS*
2000 (paper 2002) 168 pp paper $16 95 cloth $29 95

*New in paperback!*
**THE TAMING OF THE TRUE**
NEIL TENNANT
"This is a bold book, perhaps even a brave book.. Anyone interested in the realist/anti-realist controversy, whichever side she takes, should read it, for it presents the most sophisticated defense of moderate anti-realism to date "—*Mind*
1997 (paper 2002) 488 pp , numerous diagrams paper $24 95 cloth $55 00

**THE ATROCITY PARADIGM**
*A Theory of Evil*
CLAUDIA CARD
What distinguishes evils from ordinary wrongs? Are some evils unforgivable? How should we respond to evils? Card offers a secular theory of evil—representing a compromise between classic utilitarian and stoic approaches—that responds to these and other questions
June 2002 320 pp $35 00

**BEFORE FORGIVING**
*Cautionary Views of Forgiveness in Psychotherapy*
Edited by SHARON LAMB and JEFFRIE G. MURPHY
In this volume, psychologist Sharon Lamb and philosopher Jeffrie Murphy argue that forgiveness has been accepted as a therapeutic strategy without serious, critical examination.
2002 288 pp $39 95

**THE USES OF SENSE**
*Wittgenstein's Philosophy of Language*
CHARLES TRAVIS
This book provides a novel interpretation of the ideas about language in Ludwig Wittgenstein's *Philosophical Investigations*.
1989 (paper 2001) 416 pp paper $27 95 cloth $114 00

**INCLUSION AND DEMOCRACY**
IRIS MARION YOUNG
Democratic processes of communication in a multicultural society often marginalize those whose forms of expression don't fit the norms of political discussion. The author, examines the ideals of political inclusion and exclusion and recommends ways to make democratic politics more inclusive.
2000 (paper 2002) 320 pp paper $18 95 cloth $29 95

**AGAINST EQUALITY OF OPPORTUNITY**
MATT CAVANAGH
These days most people take it for granted that equality of opportunity is part of what constitutes a fair society. Yet there is no clear sense of what equality of opportunity actually means. Matt Cavanagh offers a highly provocative and original new way for us to think about equality, opportunity, and the way our society is organized.
April 2002 228 pp $35 00

OXFORD

## ARTHUR C. DANTO: ART, ACTION, HISTORY

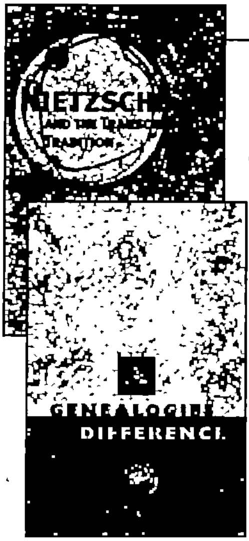
October 3-5, 2002

Miller Theater, Columbia University

The Trustees and Editors of the *Journal of Philosophy*, along with Columbia University's Philosophy Department and School for the Arts, are pleased to announce a conference to honor the work of Arthur C. Danto—as a philosopher, editor, colleague, and art critic. The conference will be held October 3-5, 2002, at Miller Theater, Columbia University, New York City, and will be open to the public.

The speakers will include: Frank Ankersmit, Hans Belting, David Carrier, Stanley Cavell, Donald Davidson, Daniel Dennett, Lydia Goehr, Daniel Herwitz, Gregg Horowitz, Philip Kitcher, Rosalind Krauss, Robert Storr, and Richard Wollheim. There will be four philosophy panels, focusing on the philosophy of art, history, and action, plus two sessions on art history. Finally, there will be a panel discussion on Saturday evening with several artists, curators, and critics addressing Danto's writings and contemporary art.

Please visit our website (www.journalofphilosophy.org or www.jphil.org) for the full program and other related information.



New from ILLINOIS

**Nietzsche and the Transcendental Tradition**

**MICHAEL STEVEN GREEN**

"A very powerful and philosophically sophisticated presentation of a position in the interpretation that has never to my knowledge had such an able advocate. . . . [Green] shows that there is more than one might think to the version of Nietzsche that has been popularized by the post-structuralists."— Richard Schacht, author of *Nietzsche* and *Making Sense of Nietzsche*

**International Nietzsche Studies**

HARDCOVER, $29 95

**Genealogies of Difference**

**NATHAN WIDDER**

"[Widder] seeks to show that the rejection of metaphysical foundations lead neither to a postmodern ironism or skepticism nor to a negative theology. . . . A pertinent and instructive contribution to contemporary thinking in the area of philosophy and political theory." — Keith Ansell-Pearson, author of *Germinal Life: The Difference and Repetition of Deleuze*

HARDCOVER, $29 95

*available from your bookseller, or visit* www.press.uILLINOIS.edu

# THE JOURNAL OF PHILOSOPHY

18\10\02

VOLUME XCIX, NUMBER 9
SEPTEMBER 2002

*page*
445 *Open and Closed Impartiality* Amartya Sen
470 *Is There a Valid Experimental Argument for Scientific Realism?* Peter Achinstein

496 NEW BOOKS: PAPERBACK REPRINTS
498 NOTES AND NEWS

*Published by* **The Journal of Philosophy, Inc.**

# THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

*Editors*: Brian Barry, Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.
You may also visit our website at: www.journalofphilosophy.org

# THE JOURNAL OF PHILOSOPHY

2002

SUBSCRIPTIONS (12 issues)

| | |
|---|---|
| Individuals | $35.00 |
| Libraries and Institutions | $75.00 |
| Students, retired/unemployed philosophers | $20.00 |
| Postage outside the U.S. | $15.00 |

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Ephrata, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by writing to us.

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.
POSTMASTER: Send address changes to the *Journal of Philosophy* at MC 4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2002 by the Journal of Philosophy, Inc. ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY

VOLUME XCIX, NO. 9, SEPTEMBER 2002

18|10|02

# OPEN AND CLOSED IMPARTIALITY*

The role of impartiality in the evaluation of social judgments and societal arrangements is well recognized in moral and political philosophy. Here, I want to discuss a basic distinction between two quite different ways of invoking impartiality. The procedures involve disparate interpretations of the demands of impartiality, and can correspondingly have rather dissimilar substantive implications. The two approaches will be called *closed* and *open* impartiality, respectively. The distinction turns on whether or not the exercise of impartial assessment is confined (or more accurately, attempted to be confined) to a fixed group: I shall call it the *focal group*.

With *closed impartiality*, the procedure of making impartial judgments invokes only the members of the focal group itself. For example, the Rawlsian method of "justice as fairness" uses the device of an "original contract" between the citizens of a given polity.[1] No outsider is involved in such a contractarian procedure, or is a party to the original contract (neither directly nor through representatives). For the members of the focal group, the "veil of ignorance" requires them to be ignorant of their exact identity *within* the focal group, and this can be an effective procedure for overcoming individual partialities within the focal group. But even under the veil of ignorance, a

---

* This essay draws on my Wessons Lectures ("Democracy and Social Justice") at Stanford University, Ethics Program, given on January 15-16, 2001. I have greatly benefited from discussions with Akeel Bilgrami and Thomas Scanlon, and also from the comments of Kenneth Arrow, G.A. Cohen, John Deigh, Nick Denyer, Barbara Fried, Isaac Levi, Mozaffar Qizilbash, Emma Rothschild, Debra Satz, Patrick Suppes, Kotaro Suzumura, and Bernard Williams.

[1] John Rawls, *A Theory of Justice* (Cambridge: Harvard, 1971)—hereafter TJ; and *Political Liberalism* (New York: Columbia, 1993)—hereafter PL.

0022-362X/02/9909/445-69

person does know that she belongs to the focal group (and is not someone outside it), and there is no insistence at all that perspectives from outside the focal group be invoked. As a device of structured political analysis, the procedure is not geared to addressing the need to overcome group prejudices.

In contrast, in the case of *open impartiality*, the procedure of making impartial judgments can (and in some cases, must) invoke judgments inter alia from outside the focal group. For example, in Adam Smith's[2] use of the device of "the impartial spectator," the demands of impartiality require the invoking of disinterested judgments of "any fair and impartial spectator," not necessarily (indeed sometimes ideally not) belonging to the focal group. Both approaches demand impartiality, but through different procedures, which can substantially influence the reach as well as the results of the respective methods.

The contrast applies to individual decisions as well as group choices. In the individual context, open impartiality is aimed at removing the biasing influences of one's objectives or interests or prejudices, not only in contrast with those of others in some group of which one is a member (there can, of course, be many such groups related, for example, to nationality, class, profession, and so on), but also of nonmembers of each group. Open impartiality is a demand that the viewpoints of others, whether or not belonging to some group of which one is specifically a member, receive adequate attention.

Smith's strategy in invoking impartial spectators makes room for taking note of perspectives that are not overwhelmed by the biases of one's own groups. The partialities of group-based thinking are to be identified, with an attempt to override them. Given the powerful influence, which can be implicit and not always clearly perceived, of one's situation and group affiliations, it is necessary to examine the perspectives of differently situated spectators—from far and near—to overcome partiality in general.[3] In contrast, the "veil of ignorance" as applied within a given society or polity provides a procedure for a more limited—*intra*societal or *intra*polity—impartiality (with the further presumption of seeing an individual primarily as a member of some uniquely defined group, such as a nation or a "people"). Whereas impartiality, broadly conceived, requires the removal of different types of biasing influences, each model of closed impartiality is devised to address one possible source of bias—in the form of interpersonal variations within a particular group.

---

[2] *The Theory of Moral Sentiments* (1759; revised edition, 1790; republished, New York: Oxford, 1976); hereafter TMS.

[3] See TMS, Book III. As will be discussed, the Smithian procedure substantially differs also from the procedure, often suggested as a "corrective" to the Rawlsian scheme, of invoking one gigantic contractarian exercise involving all the people of the world, sometimes called the "cosmopolitan" or "global" version of justice as fairness—on this, see Thomas W. Pogge, *Realizing Rawls* (Ithaca: Cornell, 1989).

## I. THE MAIN THEMES

I shall examine some of the differences between the two approaches to impartiality and their implications. Among other things, I shall argue that closed impartiality has some very specific limitations. The difficulties from which closed impartiality suffers include the following.

> (1) *Procedural parochialism*: closed impartiality is devised to eliminate partiality toward the vested interests or personal objectives of individuals in the focal group, but it is not designed to address the limitations of partiality toward the shared prejudices or biases of the focal group itself.

Procedural parochialism may or may not, in fact, be seen as a problem. In some approaches to social judgments, there is no particular interest in avoiding group leanings—indeed, sometimes quite the contrary. To illustrate, some versions of communitarianism may even celebrate the "local" nature of such priorities. The same may apply to other forms of local justice. To take an extreme case, when the Taliban rulers of Afghanistan insisted that Osama bin Laden should be tried only by a group of Islamic clerics, all committed to the Shariah, the need for some kind of impartiality (against offering personal favors or partial treatment to bin Laden) was not denied—at least not in principle.[4] Rather, what was being proposed was that the impartial judgments should come from a closed group of people who all accepted a particular religious and ethical code. There is, thus, no tension in such cases between closed impartiality and the underlying affiliative norms.

In other cases, however, when a procedure of closed impartiality is combined, in one way or another, with universalist intentions, procedural parochialism may well be seen as a difficulty. This, I shall argue, is the case with Rawlsian "justice as fairness." Despite the thoroughly nonparochial intentions of the general Rawlsian approach, the use of closed impartiality involved in the "original position" (with its program of impartial assessment confined only to members of the focal group under a "veil of ignorance" regarding individual interests and goals) does not, in fact, include any procedural guarantee against being swayed by local group prejudices.

---

[4] The reference here is only to the principles of justice that the Taliban rulers were invoking—not to their practice.

> (2) *Inclusionary incoherence*: inconsistencies can potentially arise in the exercise of "closing" the group when the decisions to be taken by any focal group can influence the size or composition of the group itself.

For example, when the size or composition of the population of a country (or a polity) is itself influenced—directly or indirectly—by the decisions taken in the original position (in particular, the choice of the basic social structure), the focal group would vary with decisions that are meant to be taken by the focal group itself. Structural arrangements, such as "the difference principle," cannot but influence the pattern of social—and biological—intercourse and thus generate different populations. This is not to deny the possible existence of a "fixed point" (with suitable assumptions regarding continuity) such that the decisions of a given focal group lead exactly back to the same focal group. But the problem of possible inconsistency cannot be ruled out in general when decisions to be taken by a focal group influence the composition of the focal group itself.

> (3) *Exclusionary neglect*: closed impartiality can exclude the voice of people who do not belong to the focal group, but whose lives are affected by the decisions of the focal group. The problem is not adequately resolved by multistaged formulations of closed impartiality, as in Rawls's "the law of peoples."

This problem will not arise if decisions taken by the focal group (for example, in the original position) do not have any effect on anyone outside the focal group. That would be quite extraordinary unless people live in a world of completely separated communities. This issue can be particularly problematic for "justice as fairness" in dealing with justice across borders, since the basic social structure chosen for a society can have an influence on the lives not only of members of that society, but also of others (who are not accommodated in the original position for that society). There can be much vexation without representation.

As it happens, the first two problems (namely, procedural parochialism and inclusionary incoherence) have hardly received any systematic attention at all—not even an adequate identification. The third problem (exclusionary neglect), in contrast, has received much attention already, in one way or another.[5] John Rawls[6] himself has addressed this issue, specifically in the context of justice across borders through his proposal of "the law of peoples," which invokes a second original position between representatives of different polities (or "peoples"). In this context, we have to see clearly why the demands of "global justice" may substantially differ from those of "international justice."[7] I shall argue that the Smithian device of the impartial spectator has insights to offer on this difficult subject.

In the next section, Smith's defense of open impartiality is briefly examined. In section III, I consider Rawls's critique of Smith. Section IV investigates the complementarity between Smithian and Rawlsian reasoning, and draws on Rawls's analysis of objectivity in moral and political philosophy, and the particular role of a "public framework of thought," which can take us beyond the restricted structure of "justice as fairness." In sections V-VII, the three identified problems of closed impartiality are scrutinized.

## II. ADAM SMITH AND THE IMPARTIAL SPECTATOR

The Smithian device of the "impartial spectator" differs substantially from the closed impartiality of "justice as fairness." The basic idea is pithily put by Smith in *The Theory of Moral Sentiments*, as the requirement, when judging one's own conduct, to "examine it as we imagine an impartial spectator would examine it," or as he elaborated in a later edition of the same book: "to examine our own conduct as we imagine any other fair and impartial spectator would examine it."[8]

The insistence on impartiality in modern moral and political philosophy does, of course, have strongly Kantian influence. Even though Smith's exposition of this idea is less remembered, there are substantial points of similarity between the Kantian and Smithian approaches. In fact, Smith's analysis of "the impartial spectator" has some claim to being the pioneering idea in this general enterprise of interpreting impartiality and of formulating the demands of fairness, which so engaged the world of European enlightenment. Smith's ideas were not only influential among such "enlightenment thinkers" as Condorcet (who was also a pioneering social-choice theorist),[9] but Immanuel Kant, too, knew *The Theory of Moral Sentiments* (originally published in 1759), and commented on it in a letter to Markus Herz in 1771 (somewhat earlier than his *Groundwork*, 1785, and *Critique of Practical Reason*, 1788).[10]

---

[5] See, for example, Charles R. Beitz, *Political Theory and International Relations* (Princeton: University Press, 1979); also Pogge; and Brian Barry, *Theories of Justice, Volume I* (Berkeley: California UP, 1989).

[6] *The Law of Peoples* (Cambridge: Harvard, 1999).

[7] I have tried to identify these issues in "Global Justice: Beyond International Equity," in Inga Kaul, I. Grunberg and M.A. Stern, eds., *Global Public Goods: International Cooperation in the 21st Century* (New York: Oxford, 1999), pp. 116-25, and also in "Justice across Borders," in *Global Justice and Transnational Politics*, Pablo De Greiff and Ciaran Cronin, eds. (Cambridge: MIT, 2002), pp. 37-51, originally presented as a lecture for the Centennial Year Celebrations of the De Paul University in Chicago in September 1998.

[8] TMS, III.1.2; the extended version occurs in the sixth edition. On the points of emphasis, see the discussion in D.D. Raphael, "The Impartial Spectator," in Andrew S. Skinner and Thomas Wilson, eds., *Essays on Adam Smith* (New York: Oxford, 1975), pp. 88-90. On the centrality of these issues in the enlightenment perspectives, particularly in the works of Smith and Condorcet, see Emma Rothschild, *Economic Sentiments: Smith, Condorcet and the Enlightenment* (Cambridge: Harvard, 2001).

There is, however, something of a parting of the ways between the Smithian approach of "the impartial spectator" and the contractarian approach, of which Rawlsian "justice as fairness" is the pre-eminent application. The need to invoke how things would look to "any other fair and impartial spectator" is a requirement that can bring in judgments that would be made by disinterested people from other societies as well. The judgmental basis of Smith's concept of fairness is, in this sense, more universal than that of a polity-prioritized theory of justice, with closed impartiality. The institutionally constructive character of the Rawlsian system also restricts the extent to which "outsiders" can be accommodated within the exercise of impartial assessment.

Even though Smith often refers to the impartial spectator as "the man within the breast," one of the main motivations of Smith's intellectual strategy is to question the soundness of reasoning alienated from the way things look to others—far as well as near. He puts the issue thus:

> In solitude, we are apt to feel too strongly whatever relates to ourselves.... The conversation of a friend brings us to a better, that of a stranger to a still better temper. The man within the breast, the abstract and ideal spectator of our sentiments and conduct, requires often to be awakened and put in mind of his duty, by the presence of the real spectator; and it is always from that spectator, from whom we can expect the least sympathy and indulgence, that we are likely to learn the most complete lesson of self-command (TMS, III.3.38, pp. 153-54).

---

[9] On the importance of Condorcet's role in social-choice theory, see Kenneth Arrow, *Individual Values and Social Choice* (New York: Wiley, 1951; extended edition, 1963). I discuss the contribution of Condorcet and his influence on modern social-choice theory (initiated by Arrow), in my Nobel lecture, "The Possibility of Social Choice," *American Economic Review*, xcix (June 1999): 349-78.

[10] See Raphael and Macfie, "Introduction," in Smith, TMS (republished edition, 1976), p. 31. It is, I suppose, immaterial that Kant chose to refer to the proud Scotsman as "the Englishman Smith."

This way of invoking impartiality, including "real spectators" (distinct from social contractors), has a very different reach from the impartiality in the closed form of contractarian reasoning.

Smith requires the impartial spectator to go beyond reasoning that may—perhaps imperceptibly—be constrained by local conventions of thought, and to examine deliberately, as a procedure, what the accepted conventions would look like from the perspective of a "real spectator" at a distance. This issue, as will be discussed presently, is particularly relevant to the challenge of "procedural parochialism." Smith's justification of such a procedure of open impartiality is spelled out thus:

> We can never survey our own sentiments and motives, we can never form any judgment concerning them; unless we remove ourselves, as it were, from our own natural station, and endeavour to view them as at a certain distance from us. But we can do this in no other way than by endeavouring to view them with the eyes of other people, or as other people are likely to view them (TMS, III.1.2, p. 110).

Smithian reasoning, thus, not only admits, but also requires, the use of impartial spectators at a distance, and the procedure of impartiality to be used is open and broad, rather than closed and restricted.

## III. ON RAWLS'S INTERPRETATION OF SMITH

The foregoing argument indicates that there can be substantial differences between the open impartiality of the impartial spectator and the closed impartiality of the social contract. But can the impartial spectator really be the basis of a viable approach to moral or political assessment without being, directly or indirectly, parasitic on some version of closed impartiality, such as contractarianism? As it happens, this issue has been addressed by Rawls himself in the *Theory of Justice*, when he comments on the general device of the impartial spectator (TJ, 183-92). Rawls interprets the impartial spectator conception as one particular example of "the ideal observer" approach (TJ, 184). Seen in this way, the idea allows some freedom, as Rawls rightly notes, about how we may proceed from there to make the conception more specific. He argues that interpreted in this way, "there is no conflict so far between this definition and justice as fairness" (TJ, 184). Indeed, it "may well be the case that an ideally rational and impartial spectator would approve of a social system if and only if it satisfies the principles of justice which would be adopted in the contract scheme" (TJ, 184-85).

This is certainly a possible interpretation of an "ideal observer," but it is definitely not, as we have already seen, Smith's conception of "the impartial spectator." It is indeed the case that the spectator can take note of what may be expected had there been such a contract, but Smith requires the impartial spectator to go beyond that and at least see what the issues would look like with "the eyes of other people"—from the perspective of "real spectators" far and near.

Indeed, Rawls too goes on to note that "while it is possible to supplement the impartial spectator definition with the contract point of view, there are other ways of giving it a deductive basis" (TJ, 185). Rawls then proceeds, however, by looking at David Hume's rather than Smith's writings. This, not surprisingly, leads Rawls to consider the alternative of making the impartial spectator rely on "satisfactions" generated by sympathetic consideration of the experiences of others: "the strength of his approval is determined by the balance of satisfactions to which he had sympathetically responded" (TJ, 186). This, in turn, takes Rawls to the interpretation that the impartial spectator may really be a "classical utilitarian" in disguise. Once that extremely odd diagnosis is made, Rawls's response is, of course, quite predictable—and predictably forceful. He points out that already in the first chapter of the *Theory of Justice*, he has discussed why "there is a sense in which classical utilitarianism fails to take seriously the distinction between persons" (TJ, 187).

Indeed, in discussing the history of classical utilitarianism, Rawls lists Smith among its early proponents, along with Hume (TJ, 22-23, footnote 9). This, of course, is just a wrong diagnosis, since Smith had firmly rejected the utilitarian proposal of basing ideas of the good and the right on pleasure and pain, and had also spurned the view that the reasoning needed for complex moral judgments can be reduced simply to counting pleasure and pain. Indeed, more generally, Smith was critical even of looking for some one thing as a simpliciter of our diverse concerns and priorities:

> By running up all the different virtues to this one species of propriety, Epicurus indulged a propensity, which is natural to all men, but which philosophers in particular are apt to cultivate with a peculiar fondness, as the great means of displaying their ingenuity, the propensity to account for all appearances from as few principles as possible. And he, no doubt, indulged this propensity still further, when he referred all the primary objects of natural desire and aversion to the pleasures and pains of the body (TMS, VII.ii.2.14, p. 299).

Thus, the Rawlsian interpretation of Smith and of his use of the "impartial spectator" is altogether mistaken.[11] More importantly, the impartial-spectator approach need not, in fact, be based either on Rawlsian contractarianism or on Benthamite classical utilitarianism—the only two options Rawls considers. Rather, the kind of diverse moral and political concerns that he himself discusses so illuminatingly are precisely the ones that the impartial spectator too has to grapple with, but without the additional (and in the Smithian perspective, inescapably arbitrary) insistence on closed impartiality. In the approach of the impartial spectator, the need for the discipline of ethical and political reasoning firmly remains, and also the requirement of impartiality stays paramount, and it is only the "closing" of that impartiality that is absent.[12] The impartial spectator can work and enlighten without being either a social contractor or a camouflaged utilitarian.

---

[11] Given Rawls's command over the history of ideas and his extraordinary generosity in presenting the views of others, it is uncharacteristic that he pays so little

## IV. OPEN IMPARTIALITY AND RAWLSIAN REASONING

Even though the argument for giving more room to open impartiality is not a specific critique of Rawlsian justice as fairness, it is inter alia a demand for going well beyond that highly successful framework of moral and political investigation. I should emphasize, however, that the exploration of open impartiality can inter alia draw on the discipline of reasoning that we have learned from Rawls's own work.[13] For example, in the "open" conception of impartiality, there is an appeal to objectivity in some form, for which Rawls's discussion of the

---

attention to the writings of Smith, especially TMS. In Rawls's far-reaching *Lectures on the History of Moral Philosophy* (Barbara Herman, ed. (Cambridge: Harvard, 2000)), Smith does get five mentions, but these passing references are confined to (1) Smith's being a Protestant, (2) a friend of Hume, (3) an amusing user of words, (4) a successful economist, and (5) the author of the *Wealth of Nations* published in the same year (1776) in which Hume died. In general, it is interesting to see how little attention the Professor of Moral Philosophy at Glasgow, so influential in economics and in philosophical thinking of his time (including Kant's), gets from contemporary moral philosophers.

[12] I have argued elsewhere—"Consequential Evaluation and Practical Reason," this JOURNAL, XCVII, 9 (September 2000): 477-502—that there is a case for paying more attention to the Kantian idea of "imperfect obligations" (relatively neglected in contemporary moral and political philosophy) which involves broader reasoning, with some similarities with Smith's impartial spectator, instead of seeing the Kantian perspective as being mainly confined to the framework of "perfect obligations," or to contractarian constructions.

[13] I must also take this opportunity of recording my deep personal debt to Rawls, since my own understanding of moral and political philosophy (and also of welfare economics) has been deeply influenced by what I have learned from him, as is acknowledged in my *Collective Choice and Social Welfare* (San Francisco: Holden-Day, 1970; republished, Amsterdam: North-Holland, 1979); *On Economic Inequality* (New York: Oxford, 1973; expanded edition, 1997); *On Ethics and Economics* (New York: Blackwell, 1987); *Inequality Re-examined* (Cambridge: Harvard, 1992).

nature of objectivity in moral and political arguments is particularly relevant.

Smith's use of the impartial spectator relates to contractarian reasoning in a somewhat similar way to that in which models of fair *arbitration* (which can be done by anyone) relate to those of fair *negotiation* (in which participation is confined to the members of the group in the original contract). In Smithian analysis, the relevant judgments can come from outside the perspectives of the negotiating protagonists; indeed, they can come from, as Smith puts it, "any fair and impartial spectator."

There are, in fact, significant similarities between parts of Rawls's own reasoning and the exercise of fair arbitration involved in open impartiality. Despite the "contractarian" form of Rawlsian theory of justice, the social contract is not the only device that Rawls invokes in developing his theory.[14] Indeed, much of the reflective exercise happens even before the hypothetical contract is supposed to occur. The "veil of ignorance" can be seen as a meta-ethical demand of impartiality which is meant to constrain any person's moral and political reflections whether or not a contract is ultimately invoked. Furthermore, while the form of that exercise of impartiality remains "closed" in the sense already discussed, it is clear that Rawls's intentions include inter alia the elimination of the hold of arbitrary influences related to past history (as well as individual advantages). In seeing "the original position" as "a device of representation," Rawls attempts to address various types of arbitrariness which may influence our actual thinking, which have to be subjected to ethical discipline to arrive at an impartial point of view:

> The original position, with the formal features I have called 'the veil of ignorance', is this point of view.... These contingent advantages and accidental influences from the past should not affect an agreement on the principles that are to regulate the institutions of the basic structure itself from the present into the future (TJ, 23).

Indeed, given the use of the discipline of the "veil of ignorance," the parties (that is, individuals under this veil) would already agree with each other when the point comes to negotiate a contract. In fact, noting this, Rawls does ask whether a contract is at all needed given the precontract agreement. He explains that despite the agreement that would precede the contract, the original contract does have a significant role because the act of contracting—even in its hypothetical form—is itself important, and because the contemplation of the act of contracting—with a "binding vote"—may influence the precontractual deliberations that occur:

---

[14] This is, in fact, a particularly good moment to re-examine the extraordinary reach of Rawlsian reasoning, since we have been lucky enough to have had a veritable feast of recent contributions from Rawls over the last three years, which have consolidated and extended his earlier writings. See Rawls, *Collected Papers*, Samuel Freeman, ed. (Cambridge: Harvard, 1999); *The Law of Peoples* (1999); *Lectures on the History of Moral Philosophy* (2000); *A Theory of Justice* (Cambridge: Harvard, revised edition, 2000); *Justice as Fairness: A Restatement*, Erin Kelly, ed. (Cambridge: Harvard, 2001).

> Why, then, the need for an agreement when there are no differences to negotiate? The answer is that reaching a unanimous agreement without a binding vote is not the same thing as everyone's arriving at the same choice, or forming the same intention. That it is an undertaking that people are giving may similarly affect everyone's deliberations so that the agreement that results is different from the choice everyone would have otherwise made.[15]

Thus, the original contract remains important for Rawls, and yet a substantial part of Rawlsian reasoning concerns precontractarian reflections, and in some ways runs on parallel lines to Smith's procedure involving fair arbitration. What distinguishes the Rawlsian method, even in this part, from the Smithian approach is the "closed" nature of the participatory exercise that Rawls invokes through restricting the "veil of ignorance" to the members of a given focal group.

Rawls's remarks on objectivity are particularly relevant to the Smithian approach. The issue of objectivity, which is closely related to open impartiality, shares something with the often-repeated demand that justice should not only be done, but also be "seen to be done."[16] The rationale of the demand lies not just in the instrumental argument that disbelief in the fairness of judgments made by the legal establishment can make it harder to *implement* legally arrived decisions. But it also appeals, more foundationally, to the sense that, if impartial spectators—near and far—with the best of effort cannot see that a judgment is just, then even its *correctness* could be in some doubt.

In pursuing these demands of public comprehension and observability, we can get considerable guidance from the Rawlsian exposition of objectivity in ethics and political philosophy, particularly from his focus on the demand for "a public framework of thought": "we look at our society and our place in it objectively: we share a common standpoint along with others and do not make our judgments from a personal slant."[17] Judgments of justice cannot be an entirely private affair—unfathomable to others—and the Rawlsian invoking of "a public framework of thought," which does not in itself demand a "contract," is a critically important move.

---

[15] Rawls, "Reply to Alexander and Musgrave," in *Collected Papers*, pp. 232-53, here p. 249; see also Tony Laden, "Games, Fairness and Rawls's *A Theory of Justice*," *Philosophy and Public Affairs*, xx (1991): 189-222.

[16] This is one of the major themes explored in my Wessons Lectures at Stanford University, "Democracy and Social Justice" (January 2001).

That move is further consolidated by Rawls's argument, particularly in *Political Liberalism*, that the relevant standard of the objectivity of ethical principles is basically congruent with their defensibility within a public framework of thought (there is no further test):

> To say that a political conviction is objective is to say that there are reasons, specified by a reasonable and mutually recognizable political conception (satisfying those essentials), sufficient to convince all reasonable persons that it is reasonable (PL, 119).

This thesis links closely also with Thomas Scanlon's[18] argument that "thinking about right and wrong is, at the most basic level, thinking about what could be justified to others on grounds that they, if appropriately motivated could not reasonably reject" (*ibid.*, p. 5).

The accord that is sought need not, of course, be anywhere near total.[19] The concurrence need not go beyond a partial ordering with limited articulation, which can nevertheless make firm and useful statements. The agreements arrived at need not demand that some proposal is uniquely just, but perhaps only that it is plausibly just, or at least not manifestly unjust. Indeed, the demands of reasoned practice can, in one way or another, live with a good deal of incompleteness or unresolved conflicts.[20] In particular, the acknowledgment of some incompleteness does not indicate that all is lost.[21] The agreement to emerge from "a public framework of thought" can be of a correspondingly permissive kind.

---

[17] TJ, 516-17; more extensively, see TJ, 513-20, and PL, 110-16.

[18] *What We Owe to Each Other* (Cambridge: Harvard, 1998); see also his "Contractualism and Utilitarianism," in Sen and Bernard Williams, eds., *Utilitarianism and Beyond* (New York: Cambridge, 1982), pp. 103-28.

[19] As Rawls puts it, "given the many obstacles to agreement in political judgement even among reasonable persons, we will not reach agreement all the time, or perhaps even much of the time" (PL, 118).

[20] I have discussed the relevance of incompleteness and partial orderings in *Collective Choice and Social Welfare*; *On Economic Inequality*; "Internal Consistency of Choice," *Econometrica*, LXI (1993): 495-521; "Maximization and the Act of Choice," *Econometrica*, LXV (1997): 745-75; "Consequential Evaluation and Practical Reason." For a different but related approach to incompleteness, see Isaac Levi, *Hard Choices* (New York: Cambridge, 1986), and *The Covenant of Reason: Rationality and the Commitments of Thought* (New York: Cambridge, 1997).

[21] Indeed, even when an incompleteness is not just tentative (resolvable on the basis of further discussion and scrutiny), but in fact claimed to be "assertive"

It might, however, be thought that a serious difficulty can arise from elsewhere, related to a different issue, in trying to extend the Rawlsian "public framework of thought" beyond the borders of a country, or a polity. Can comprehension and thought cross geographical borders? While some are evidently tempted by the belief that we cannot follow each other across the borders of a given community, a particular nation, or a specific culture (a temptation that has been fueled particularly by the popularity of some versions of communitarian separatism), there is no particular reason to presume that interactive communication and public engagement can be sought only within such boundaries (or within the confines of those who can be seen as "one people"). Smith saw the possibility that the impartial spectator could draw on the understanding of people who are far as well as those who are near.

This was indeed a significant theme in the intellectual concerns of Enlightenment writers. While commenting on the importance of increased communication in expanding the reach of our sense of justice, Hume[22] noted:

> ...again suppose that several distinct societies maintain a kind of intercourse for mutual convenience and advantage, the boundaries of justice still grow larger, in proportion to the largeness of men's views, and the force of their mutual connexions (*ibid.*, p. 25).

The possibility of communication and cognizance across the borders should be no more absurd today than it was in Smith's eighteenth-century world. Indeed, right at this time, as the world is engaged in discussions and debates about the unacceptability—and the roots—of global terrorism, it is hard to think that we cannot understand each other across the borders of our polity.[23] Rather, it is the firmly "open" outlook, which Smith's "impartial spectator" invokes,

---

(expressing the view that the alternatives *cannot* be ordered or classified on the basis of a congruence of nonrejectable principles), that too can be an affirmative claim (not an expression of defeat). These questions are further examined in "Justice and Assertive Incompleteness," the second Rosenthal Lecture at the Northwestern University Law School (1998), and "Incompleteness and Reasoned Choice," mimeographed essay written for a festschrift for Isaac Levi (2002).

[22] *An Enquiry Concerning the Principles of Morals* (1777; republished, La Salle, IL: Open Court, 1966).

[23] Elsewhere, I have tried to discuss the role of crosscultural interactions in the history of cultures (including in the emergence of "Western civilization" and "Western science")—"Civilized Imprisonments," in *The New Republic* (June 10, 2002): 28-33.

that may be in some need of reassertion today. It can make a difference to our understanding of the demands of impartiality in moral and political philosophy in the interconnected world in which we live.

### V. PROCEDURAL PAROCHIALISM

That the veil of ignorance, as defined by Rawls, does not, as a procedure, provide a bar against catering to local group interest is easy to see. Its extensive implications will be discussed presently (in section VII). But the limitation of the closed framework of impartiality can go much further, since there is no procedural barricade here against susceptibility to local prejudices either. The Rawlsian exercise involves institutional reasoning among people "who are born into that society in which they lead their lives" (PL, 23). What is a matter of concern here is the absence of some procedural insistence on forceful scrutiny of local values that may, on further scrutiny, turn out to be preconceptions and biases that are common in a focal group.

The contrast with Smith's procedure involving the "impartial spectator" is particularly relevant here. Smith noted, in a chapter entitled "On the Influence of Custom and Fashion upon the Sentiments of Moral Approbation and Disapprobation," that "the different situations of different ages and countries are apt...to give different characters to the generality of those who live in them, and their sentiments concerning the particular degree of each quality, that is either blamable or praise-worthy, vary, according to that degree which is usual in their own country, and in their own times" (TMS, V.2.7, p. 204).[24] Indeed, it may well emerge that what is taken to be perfectly natural and normal in a society cannot survive a broad-based and less limited scrutiny.[25] Smith gives various examples of parochial thought, swayed by the hold of local mores.

> ...the murder of new-born infants was a practice allowed of in almost all the states of Greece, even among the polite and civilized Athenians; and whenever the circumstances of the parent rendered it inconvenient to bring up the child, to abandon it to hunger, or to wild beasts, was regarded without blame or censure.... Uninterrupted custom had by this time so thoroughly authorized the practice, that not only the loose maxims of the world tolerated this barbarous prerogative, but even the

---

[24] Smith noted that the "effects of custom and fashion...upon moral sentiments of mankind" frequently reflect the variability of circumstances that demand different priorities, so that we cannot, in general, "complain that the moral sentiments of men are very grossly perverted" (p. 209). But he also pointed out that sometimes local values are in very great need of a more open scrutiny.

[25] The critical role of comparative reference and scrutiny supplements the ethical inclusiveness of a nonparochial approach, on which see Martha Nussbaum with Respondents, *For Love of Country* (Boston: Beacon, 1996).

> doctrine of philosophers, which ought to have been more just and accurate, was led away by the established custom, and upon this, as upon many other occasions, instead of censuring, supported the horrible abuse, by far-fetched considerations of public utility. Aristotle talks of it as of what the magistrates ought upon many occasions to encourage. Plato is of the same opinion, and, with all that love of mankind which seems to animate all his writings, no where marks this practice with disapprobation (TMS, V.2.15, p. 210).

Smith's insistence that we must inter alia view our sentiments from "a certain distance from us" is, thus, motivated by the object of scrutinizing not only the influence of vested interest, but also the impact of entrenched tradition and custom. While Smith's example of infanticide remains sadly relevant in some societies even today, many of his other examples have relevance to other contemporary societies. This applies, for example, to Smith's insistence that "the eyes of the rest of mankind" must be invoked to understand whether "a punishment appears equitable."[26] Scrutiny from a "distance" may be useful for practices as different as the stoning of adulterous women in Taliban's Afghanistan, selective abortion of female fetuses in China, Korea, and parts of India,[27] and the use of capital punishment (with or without opportunity for celebratory public jubilation) in the United States. An impartial assessment requires not only the avoidance of the impact of individual vested interests, but also an exacting scrutiny of parochial moral and social sentiments, which may influence the ideas and outcomes in locally separated "original positions."

In the part of Rawlsian analysis which relates to the importance of a "public framework of thought," and the need to "look at our society and our place in it objectively" (TJ, 516-17) there is, in fact, much in common with Smithian reasoning. And yet the procedure of segregated "original positions," operating in devised isolation, is not conducive to guaranteeing an adequately objective scrutiny of social conventions and parochial sentiments, which may influence what rules are chosen in the original position. When Rawls says that "our moral principles and convictions are objective to the extent that they have been arrived at and tested by assuming [a] general standpoint," he is attempting to unlock the door for an open scrutiny, and yet later on in the same sentence, the door is partially bolted by the procedural form of the requirement: "and by assessing the arguments for them by the restrictions expressed by the conception of the original position" (TJ, 517).

---

[26] Smith, *Lectures on Jurisprudence*, R.L. Meek, D.D. Raphael and P.G. Stein, eds. (New York: Oxford, 1978; reprinted, Indianapolis: Liberty, 1982), p. 104.

[27] See my "The Many Faces of Gender Inequality," *The New Republic* (September 17, 2001): 35-40; longer version, *Frontline* (November 2001): 4-14.

Rawls insists on the closed nature of the "original position":

> ...I assume that the basic structure is that of a closed society: that is, we are to regard it as self-contained and as having no relations with other societies.... That a society is closed is a considerable abstraction, justified only because it enables us to focus on certain main questions free from distracting details (PL, 12).

Rawls goes on to note one limitation of this approach: "At some point a political conception of justice must address the just relations between peoples, or the law of peoples, as I shall say." That issue is indeed addressed by Rawls's later work (*The Law of Peoples*), and will be examined later on in this essay (in section VII). But the "just relations between peoples" is an altogether different issue from the need for an open scrutiny, through a nonparochial procedure, of the values and practices of any given society or polity. The closed formulation of the program of Rawlsian "original position" extracts a heavy price in the absence of any procedural guarantee that local values will be subjected to an open scrutiny, and this loss is more than that of a "distracting detail."

The Rawlsian "veil of ignorance" in the "original position" is a very effective device in making people see beyond their personal vested interests and goals. And yet it may do little to ensure an open scrutiny of local and possibly parochial values. There is something to learn from Smith's skepticism about the possibility of going beyond local presuppositions—or even implicit bigotry—"unless we remove ourselves, as it were, from our own natural station, and endeavour to view them as at a certain distance from us." The Smithian procedure includes, as a result, the insistence that the exercise of impartiality must be open (rather than locally closed), since "we can do this in no other way than by endeavouring to view them with the eyes of other people, or as other people are likely to view them" (TMS, III.1.2, p. 110). Even as the Rawlsian "veil of ignorance" addresses effectively the need to remove the influence of vested interests and personal slants of the diverse individuals within the focal group, it abstains from invoking the scrutiny of (in Smith's language) "the eyes of the rest of mankind." Something more than an "identity blackout" *within* the confines of the local focal group would be needed to address this problem. It is in this sense that the procedural device of closed impartiality in "justice as fairness" can be seen as parochial (despite Rawls's ecumenical intentions).

## VI. INCLUSIONARY INCOHERENCE AND FOCAL GROUP PLASTICITY

The fact that the members of the focal group have a status in the contractarian exercise that nonmembers do not enjoy creates problems even when we confine our attention to one society—or one "people"—only. The size and composition of the population may alter with public policies (whether or not they are dedicated "population policies") and the populations can vary even with the "basic structure" of the society. Any rearrangement of economic, political, or social institutions, such as the "basic structure of the society" (including such rules as the "difference principle") would tend to influence the size and composition of the group that would be born, through changes in marriages, mating, cohabitation and other parameters of reproduction.[28] The focal group that would be involved in the choice of the "basic structure" would be influenced by that choice itself, and this makes the "closing" of the group for closed impartiality a potentially incoherent exercise.[29]

To illustrate this problem of group plasticity, suppose there are two institutional structures *A* and *B* that would yield, respectively, five million and six million people. Who, we have to ask, are included in the *original position* (OP) in which social decisions are made which would inter alia choose between *A* and *B* and thus influence the size and composition of the respective population groups? Suppose we take the larger group of six million people as the focal group who are included in the OP, and suppose also that it turns out that the institutional structure chosen in the corresponding OP is *A*, leading to an actual population of five million people. But, then, the focal group was wrongly specified. (We may also ask: How did the nonexistent—indeed *never* existent—extra one million people participate in the OP?) If, on the other hand, the focal group is taken to be the smaller number of five million people, what if the institutional structure chosen in the corresponding OP is *B*, leading to an actual population of six million people? Again, the focal group would turn out to be wrongly specified. Also, the additional one million people, then, did not participate in the OP, which would have decided the institutional structures that would extensively influence their lives (indeed not just *whether* they are to be born or not, but also *other* features of their actual lives). If the decisions taken in the OP influence the size and composition of the population, and if the population size and composition influence the nature of the OP, then there is no way of guaranteeing that the focal group associated with the OP is coherently characterized.

---

[28] On this, see Derek Parfit, *Reasons and Persons* (New York: Oxford, 1984).

[29] This issue has been discussed in my "Normative Evaluation and Legal Analogues," presented at a conference on "Norms and the Law," at the School of Law of Washington University in St. Louis (March 2001).

The foregoing difficulty applies even when we consider the so-called "cosmopolitan" or "global" version of the Rawlsian "justice as fairness," including all the people in the world in one large contractual exercise. The population plasticity problem is not dependent on societal specification, and would apply no matter whether we consider one "people" among many peoples or all the people put together.

When the Rawlsian system is applied to one particular "people" in a larger world, however, then there are further problems. In fact, the dependence of births and deaths on the basic social structure has some parallel also in the influence of that structure on the movements of people from one country to another. This general concern has some similarity with one of Hume's[30] grounds for skepticism of the conceptual relevance as well as historical force of "the original contract":

> The face of the earth is continually changing, by the increase in small kingdoms into great empires, by the dissolution of great empires into smaller kingdoms, by the planting of colonies, by the migration of tribes.... Where is the mutual agreement or voluntary association so much talked of (*ibid.*, p. 279)?

But the point at issue, in the present context, is not only—indeed not primarily—that the size and composition of the population is continually changing, but that these changes are not independent of the basic social structures that are meant to be arrived at, in contractual reasoning, precisely in the original position itself.

We must, however, examine further whether the dependence of the focal group on the basic social structure is really a problem for Rawlsian justice as fairness. Does the focal group actually have to determine the basic social structure through the corresponding OP? The answer, of course, is straightforwardly "yes," if the parties to the OP are meant to be exactly the focal group (namely, all—and only—the members of the polity or society). But sometimes Rawls speaks of "the original position" as "simply a device of representation."[31] It might, thus, be tempting to argue that we do not have to assume that everyone in the society or polity—no more and no less—has to be a party to the original contract, and it could be argued that, therefore, the dependence of the focal groups on the decisions taken in the OP need not be a problem.

---

[30] "On the Original Contract," republished in Hume, *Selected Essays*, Stephen Copley and Andrew Edgar, eds. (New York: Oxford, 1996), pp. 274-91.

[31] "Justice as Fairness: Political Not Metaphysical," in *Collected Papers*, p. 401.

I do not think that this is an adequate rebuttal of the problem of inclusionary incoherence for two distinct reasons. First, Rawls's use of the idea of "representation" does not, in fact, amount to marshaling a wholly new set of people (or phantoms) as parties to the OP, different from the actual people in that polity. Rather, it is the *same* people under the "veil of ignorance" who are seen as "representing" themselves (but from behind "the veil"). Rawls explains this in the following way:

> This is expressed figuratively by saying that the parties are behind a veil of ignorance. In sum, the original position is simply a device of representation (*ibid.*, p. 401).[32]

Indeed, Rawls's justification of the need for a contract, which invokes (as was noted earlier) "the undertaking that people are giving,"[33] indicates concrete participation (albeit under the veil of ignorance) of the people involved in the original contract.

Second, even if the representatives were to be different people (or imagined phantoms), they would have to represent the focal group of people (for example, through the veil of ignorance of possibly being any member of the focal group). So the variability of the focal group would be reflected in the variability of the people whom the representatives represent in the original position.[34]

This would not be much of a problem (1) *if* the size of the population did not make any difference to the way the basic structure of the society may be organized (complete scale invariance), and (2) *if* every group of individuals were exactly like every other in terms of their priorities and values (complete value invariance). Neither is easy to assume without further restrictions in the structure of any substantive theory of justice.[35] Group plasticity, thus, does remain a problem for the exercise of closed impartiality, applied to a *given* focal group of individuals.

---

[32] As Rawls puts it elsewhere (PL, 23): "Justice as fairness recasts the doctrine of the social contract. ...the fair terms of social cooperation are conceived as agreed to by those engaged in it, that is, by free and equal citizens who are born into the society in which they lead their lives."

[33] "Reply to Alexander and Musgrave," in *Collected Papers*, p. 249.

[34] To forestall a possible line of response, I should emphasize that this is not a similar problem at all to the difficulty of representing members of the future generation (seen as a *fixed group*). There is, to be sure, a problem there too (for example, about how much can be assumed about the future generations' reasoning since they are not here yet), but it is nevertheless a different issue. There is a distinction between (i) the problem of what can be presumed about the agreement of the future generations (seen as a fixed group) to be represented, and (ii) the impossibility of having a fixed group to be represented, in choosing the basic structure of the society when the set of actual persons must *vary* depending on the choice of that structure.

We must, however, also ask whether the Smithian approach of the impartial spectator is not similarly troubled by incongruity arising from group plasticity, and if not, why not. It is not similarly troubled precisely for the reason that the impartial spectator does not come from a given focal group. Indeed, Smith's "abstract and ideal spectator" is a "spectator" and not a "participant" in any exercise like a group-based contract. There is no contracting group, and there is no insistence even that the evaluators (or arbitrators) must be congruent with the affected group. Even though there remains the very difficult problem of how an impartial spectator would go about deciding on such issues as variable population size (an ethical issue of profound complexity),[36] the problem of incoherence and incongruity in "inclusionary closure" in the contractarian exercise does not have an immediate analogue in the case of the impartial spectator.

## VII EXCLUSIONARY NEGLECT AND GLOBAL JUSTICE

I turn now to the problem of exclusionary neglect. Since the decisions taken about the basic structure of a society $S$ (and what results from it) can affect the lives of people in other societies, such as $T$, there is a clear asymmetry in the exercise of closed impartiality involving only the citizens of $S$, with no role being given to people in $T$. It is not hard to see a lacuna there. Indeed, problems created by this asymmetric formulation of justice have rightly received reasoned attention.[37]

---

[35] It is important also to avoid a misunderstanding, which I have already encountered in trying to present this article, that takes the form of arguing that differing populations cannot make any difference to the Rawlsian OP, since every individual is exactly like any other under the "veil of ignorance." The point to note is that even though the "veil of ignorance" makes *different individuals within a given group* ignorant of their respective interests and values (making everyone much the same in the *as if* deliberative exercise for a given group), it does not, by itself, have any implication whatever in making *different groups of individuals* have exactly the same configuration of interests and the same cluster of values. More generally, to make the exercise of closed impartiality fully independent of the size and composition of the focal group, the substantive reach of that exercise has to be severely impoverished.

[36] The complexity would have been even greater if it were necessary that these judgments must take the form of *complete* orderings, but, as has been already discussed, this is not needed for a useful public framework of thought, nor for the making of public choices based on "maximality" (on which see my "Maximization and the Act of Choice," *Econometrica*, LXV (1997): 745-75).

Rawls himself has addressed this asymmetry by invoking another "original position," this time involving representatives of different "peoples."[38] With some oversimplification—not central in the present context—the two "original positions" can be seen as being respectively *intranational* (between individuals in a nation) and *international* (between representatives of different nations). Each exercise is one of closed impartiality, but the two together cover the entire world population.[39]

The idea of one global exercise of open impartiality, treating everyone symmetrically, may appear to be deeply "unrealistic" in the way the world is organized today. Certainly, there is an institutional lacuna here. We can nevertheless invoke the insights and instructions generated by a cross-border "public framework of thought," as Smith (among many others) have done. I shall not pursue further this issue here beyond noting that the ethical relevance and practical influence of global discussions are not totally conditional on the existence of a well-organized planetary forum for gigantic institutional agreements.[40]

More immediately, even in the politically divisive world in which we live, we have to give fuller recognition to the fact that different persons across borders need not operate only through international (or "inter-people") relations. The world is divisive all right, but it is diversely divisive, and the partitioning of the global population into distinct "nations" or "peoples" is not the only line of division.[41] Nor does the national partitioning have any pre-eminent priority over other categorizations (as implicitly presumed in "the law of peoples").[42]

---

[37] See, for example, Beitz; Pogge; and Pogge, ed., *Global Justice* (Malden, MA: Blackwell, 2001).

[38] "The Law of Peoples," in Stephen Shute and Susan Hurley, eds., *On Human Rights* (New York: Basic, 1993), pp. 41-82, and *The Law of Peoples*.

[39] This does not, of course, eliminate the asymmetry between different groups of affected people, since the different polities are diversely endowed in assets and opportunities, and there would be a clear contrast between covering the world population through a sequence of prioritized impartialities, as opposed to covering it through one comprehensive exercise of impartiality (as in the "cosmopolitan" version of the Rawlsian OP).

[40] On this, see my "Consequential Reasoning and Practical Reason."

[41] It is interesting that the priority of one specific partitioning of the global population has been proposed in disparate political discussions, giving the pride of place, respectively, to various *different* single categorizations. The categorization underlying the so-called "clash of civilizations" is an example of a rival partitioning (see Samuel P. Huntington, *The Clash of Civilizations and the Remaking of the World Order* (New York: Simon and Schuster, 1996)), since national or polity-based categories do not coincide with categories of culture or civilization. The co-existence of these rival claims in itself illustrates why none of these putatively foundational partitionings can easily drown the competing relevance of the others.

Interpersonal relations across country borders go far beyond international interactions in many different ways. The "original position" of nations or "peoples" would be peculiarly restricted in dealing with many of the cross-border effects of human action. If the effects of the operation of transnational corporations are to be assessed or scrutinized, they have to be seen as what they are, namely, corporations that operate without borders, taking business decisions about registration, tax homes, and similar contingent matters according to the convenience of business. They can hardly be fitted into the model of one "people" impacting on another.

Similarly, the ties that bind human beings in relations of duty and concern across borders need not operate *through* the collectivities of the respective nations. To illustrate, a feminist activist in America who wants to do something to remedy particular features of women's disadvantage in, say, Sudan would tend to draw on a sense of affinity that need not work through the sympathies of the American nation for the predicament of the Sudanese nation. Her identity as a fellow woman, or as a person (male or female) moved by feminist concerns, may be more important in a particular context than her citizenship, and the feminist perspective may well be introduced in an exercise of "open impartiality" without its being "subsequent" to national identities. Other identities, which may be particularly invoked in other exercises of "open impartiality," may involve class, language, literature, profession, and so on, and can provide different and competing perspectives to the priority of nation-based politics.

Even the identity of being a human being—perhaps our most basic identity—may have the effect, when fully seized, of broadening our viewpoint correspondingly. The imperatives that we may associate with our humanity may not be mediated by our membership of smaller collectivities such as specific "peoples" or "nations." Indeed, the normative demands of being guided by "humanity" or "humaneness" can build on our membership of the wide category of human beings—irrespective of our particular nationalities, or sects, or tribal affiliations (traditional or modern).[43]

---

[42] There is a related issue of the tyranny that is imposed by the privileging of an alleged "cultural" or "racial" identity over other identities and over nonidentity-based concerns; on this, see K. Anthony Appiah and Amy Gutman, *Color Consciousness: The Political Morality of Race* (Princeton: University Press, 1996); and Susan Moller Okin, with Respondents, *Is Multiculturalism Bad for Women?* (Princeton: University Press, 1999). I also discuss this issue in my *Reason before Identity* (New York: Oxford, 1998) and "Other People," Annual British Academy Lecture 2000; shorter version published in *The New Republic* (December 18, 2000): 23-30.

Behavioral correlates of global commerce, global culture, global politics, global philanthropy, even global protests (like those, recently, on the streets of Seattle or Washington or Melbourne or Prague or Quebec or Genoa) draw on direct relations between human beings—with their own standards, and their respective inclusions and priorities related to a variety of classifications. These ethics can, of course, be supported or scrutinized or criticized in different ways, even by invoking other intergroup relations, but the intergroup relations need not be confined to—or even be led by—international relations (or by "the law of peoples"). They may involve very many diverse groups, with identities that vary from seeing oneself as a businessman or a worker, as a woman or a man, as a libertarian or a conservative or a socialist, as being poor or rich, or as a member of one professional group or another (of, say, doctors or lawyers).[44] Collectivities of many different types may be invoked. International justice—even in the finely tuned form of "the law of peoples"—is simply not adequate for global justice.

This issue has a bearing also on contemporary discussions on human rights. The notion of human rights builds on our shared humanity. These rights are not derived from the citizenship of any country, or the membership of any nation, but are presumed to be claims or entitlements of every human being. It differs, thus, from constitutionally created rights guaranteed for specified people (such as American or French citizens). For example, the human right of a person not to be tortured or subjected to terrorist attacks is affirmed independently of the country of which this person is a citizen, and also quite irrespective of what the government of that country—or any other—wants to provide or support.

---

[43] There is a related issue, which is not central to the present discussion, but which can be important in seeing the form and reach of identity-based normative arguments. When used in the form of an identity-based reasoning that draws on reciprocity or mutual relations, consideration may be taken to be owed to others precisely on the ground that they are "fellow members" of some relevant category—broad or narrow. That line of reasoning can, however, be distinguished from arguments that make no use of the *shared membership* of the subject and the object, but nevertheless invoke ethical norms (of, say, "humanity") that may be expected to guide the behavior of any human being. The identity of the protagonist may be important even when no special importance is attached to the sharing of that identity by others. I have investigated this distinction in "Other People" and I have discussed it further in my lectures on "The Future of Identity" at the Pardee Center at Boston University (2001-2002).

[44] Similarly, dedicated activists working for global nongovernment organizations (such as OXFAM, Amnesty International, Médecins sans Frontières, Human Rights Watch, and others) explicitly focus on affiliations and associations that cut across national boundaries.

In overcoming the limitations of "exclusionary neglect," use can be made of the idea of open impartiality embedded in a universalist approach—of the kind that relates closely to Smith's concept of the impartial spectator. That broad framework of impartiality makes it particularly clear why considerations of basic human rights, including the importance of safeguarding elementary civil and political liberties, need not be contingent on citizenship and nationality, and may not be institutionally dependent on a nationally derived social contract. Also, there is no need to presume a world government, or even to invoke a hypothetical global social contract. The "imperfect obligations" associated with the recognition of these human rights can be seen as falling broadly on anyone who is in a position to help.[45]

The liberating role of open impartiality allows different types unprejudiced and unbiased perspectives to be brought into consideration, and encourages us to benefit from the insights that come from differently situated impartial spectators. In scrutinizing these insights together, there may well be some common understanding that emerges forcefully, but there is no need to presume that all the differences arising from distinct perspectives can be settled similarly. As was discussed earlier (in section IV), systematic guidance to reasoned decisions can come from incomplete orderings that reflect unresolved conflicts. Indeed, as the recent literature in "social choice theory," allowing relaxed forms of outcomes (such as partial social orderings), has made clear, social judgments are not rendered useless or hopelessly problematic just because the evaluative process leaves many pairs unranked and many conflicts unsettled.[46]

For the emergence of a shared and useful understanding of many substantive issues of rights and duties (and also of rights and wrongs), there is no need to insist that we must have agreed complete orderings or universally accepted full partitionings of the just from the unjust. For example, a common resolve to fight for the abolition of famines, or genocide, or terrorism, or slavery, or untouchability, or illiteracy, or epidemics, and so on, does not require that there be a similarly extensive agreement on the appropriate formulae for inheritance rights, or income-tax schedules, or levels of minimal wages, or term limits, or copyright laws. The basic relevance of the distinct

[45] Arguments in defense of this perspective have been presented in my "Consequential Evaluation and Practical Reason."

[46] See Arrow, Sen, and Kotaro Suzumura, eds., *Social Choice Re-examined* (Amsterdam. Elsevier, 1997).

perspectives—some congruent, some divergent—of the people of the world (diversely diverse as they are) is part of the understanding that open impartiality tends to generate. There is nothing particularly defeatist in this recognition.

## VIII. A CONCLUDING REMARK

This essay has been concerned with contrasting two ways of approaching the demands of impartiality in moral and political philosophy. The procedure of closed impartiality, particularly exemplified by contractarian devices applied to closed groups, can involve a strictly partial approach to impartiality. It suffers, as a result, from a number of distinct problems, of which "exclusionary neglect" is one. This particular deficiency has already received some attention, but it can be further investigated through the Smithian device of "impartial spectators," which radically differs from the procedure of two-stage contractarianism as in "the law of peoples" (for example, intranational and international), and also from grand proposals of a single, gigantic global contract.

The closed impartiality enshrined in the contractarian approach can suffer from other serious problems as well, including "procedural parochialism" and "inclusionary incoherence." Since these limitations have not yet received much examination at all, they demand greater—and clearer—recognition, and a substantial part of this article was devoted to them, particularly to procedural parochialism. In tackling each of these problems, the alternative of open impartiality has some merit, which is not shared by closed impartiality with its parochial formulation. The impartial spectator does indeed have something to tell us—even more vocally today than in the world that Smith knew.

AMARTYA SEN

Trinity College/Cambridge

# IS THERE A VALID EXPERIMENTAL ARGUMENT FOR SCIENTIFIC REALISM?*

> Long before atoms could be detected individually, scientists deduced their existence from the way dust motes danced in droplets of liquid; atoms making up the liquid were colliding with and jostling the dust.
>
> —*New York Times*[1]

Wesley C. Salmon[2] claims that there is a valid experimental argument for scientific realism, and that one of the best examples is that provided by Jean Perrin[3] early in the twentieth century. In 1908, Perrin conducted a series of experiments on Brownian motion on the basis of which he claimed that Avogadro's number $N$, the number of molecules in a substance whose weight in grams equals its molecular weight, is approximately $6\times10^{23}$. Perrin drew the conclusion that unobservable molecules exist (*ibid.*, pp. 213-27).

By "scientific realism," Salmon means a doctrine committed at least to the claim that unobservable entities exist. (What else, if anything, scientific realism does or should entail is controversial; the question will be taken up in sections V-VII.) In seeking an argument to establish the claim that unobservables exist, Salmon writes:

> ...I decided to try an empirical approach to the philosophical problem [of scientific realism]. Since it seemed unlikely that scientists would have been moved by the kinds of arguments supplied by philosophers, I felt that some insight might be gained if we were to consider the evidence and arguments that convinced scientists of the reality of unobservable entities. Although scientists, by and large, seem committed to the existence of a variety of unobservable entities..., the existence of atoms and molecules, as the microphysical constituents of ordinary matter, is the most clear and compelling example (*op. cit.*, pp. 213-14).

Salmon proceeds to reconstruct Perrin's argument and to claim that it establishes the existence of molecules experimentally. It is a simple step to scientific realism:

---

* In memory of Wesley C. Salmon.

For very helpful questions and suggestions, I am indebted to Joseph Berkovitz, Sean Greenberg, Gregory Morgan, and Michael Williams.

[1] March 29, 2001, p. A19.

[2] *Scientific Explanation and the Causal Structure of the World* (Princeton: University Press, 1984), pp. 213-27.

[3] "Brownian Motion and Molecular Reality," reprinted in Mary Jo Nye, ed., *The Question of the Atom* (Los Angeles: Tomash, 1984), pp. 507-601 (see *Annales de Chimie et de Physique* (1909)); and *Atoms* (Woodbridge, CT: Ox Bow, 1990).

0022-362X/02/9909/470–95

Molecules exist.
Molecules are unobservable entities.
Therefore, unobservable entities exist.

Since the first step is itself the conclusion of an argument based on Perrin's experiments with Brownian motion, we seem to have an experimental argument for scientific realism. Because this argument convinced at least some antirealist scientists of the reality of molecules,[4] Salmon challenges antirealist philosophers to say why they should not follow suit.

In what follows, I want to see how antirealist philosophers respond, or could respond, to Salmon's challenge to accept an experimental argument for scientific realism. First, however, I turn to Perrin's argument itself.

## I. PERRIN'S EXPERIMENTAL ARGUMENT

A discovery was made in 1827 by the English botanist Robert Brown that small microscopic particles suspended in a liquid do not sink but exhibit rapid, haphazard motion—so-called Brownian motion. In 1908, Perrin conducted a series of experiments involving microscopic particles of gamboge (a gum resin extracted from certain Asiatic trees) in a dilute emulsion. These particles exhibited Brownian motion which was visible using a microscope. The emulsion was contained in a cylinder of known height $h$. Perrin determined the density $D$ of the material making up the particles, the density $d$ of the liquid, the mass $m$ of the particles (all of them prepared by him to be the same weight), the temperature $T$ of the liquid, and (with microscopes) the number of suspended particles per unit volume at various heights. He performed experiments with different emulsions, particles of different sizes and mass, different liquids, and different temperatures.

Using an argument that I shall note presently, he derived the following equation, which relates the quantities just cited:[5]

---

[4] For example, in 1909, Friedrich Wilhelm Ostwald, who had previously rejected atomic theory on antirealist grounds, wrote: "*I have convinced myself that we have recently come into possession of experimental proof of the discrete or grainy nature of matter, for which the atomic hypothesis had vainly been sought for centuries, even millenia.* The isolation and counting of gas ions on the one hand...and the agreement of Brownian movements with the predictions of the kinetic hypothesis on the other hand, which has been shown by a series of researchers, most completely by J. Perrin—this evidence now justifies even the most cautious scientist in speaking of the *experimental* proof of the atomistic nature of space-filling matter"—quoted in Stephen G. Brush, "A History of Random Processes," *Archive for History of the Exact Sciences*, v (1968): 1-36.

[5] In addition to the quantities mentioned above, in equation (1), $n'$ and $n$ represent the number of Brownian particles per unit volume at the upper and lower

$$(1) \quad \frac{n'}{n} = 1 - \frac{Nmg(1 - d/D)h}{RT}$$

Employing different experimental values obtained for all the quantities in equation (1) except for $N$, Perrin used this equation to determine whether Avogadro's number $N$ is really a constant, and if so, what its value is.

He arrived at equation (1) by assuming that the motions of the visible Brownian particles are caused by collisions with the molecules making up the dilute liquid in which these visible particles are suspended. Accordingly, he also assumed that the visible particles will mirror the behavior of the invisible molecules. Finally, he assumed that molecules in a dilute solution of the sort in question will behave like molecules in a gas with respect to their vertical distribution. He then derived the following formula (the law of atmospheres) that governs a volume of gas contained in a thin cylinder of unit cross-sectional area and very small elevation $h$[6]

$$(2) \quad \frac{p'}{p} = 1 - \frac{Mgh}{RT}$$

The pressure of a gas is proportional to its density, and hence, Perrin assumed, to the number of molecules per unit volume. So he replaced the ratio of pressures $p'/p$ by the ratio $n'/n$, where $n'$ and $n$ are the number of molecules per unit volume at the upper and lower levels. He also replaced $M$ by $Nm$, where $m$ is a mass of a molecule of gas and $N$ is Avogadro's number. With these substitutions, Perrin obtained

$$(3) \quad \frac{n'}{n} = 1 - \frac{Nmgh}{RT}$$

He now transformed equation (3) into equation (1) by letting $n'$ and $n$ represent the number of Brownian particles per unit volume at the upper and lower levels; and in (3) substituting the expresion $mg(1 - d/D)$—the "effective weight" of a Brownian particle—for $mg$, the weight of a molecule.[7] Strictly speaking, in equation (1), $N$ represents

---

levels; $mg$ is the weight of a particle; $R$ is the gas constant; and $N$ is Avogadro's number.

[6] In (2), $p'$ and $p$ are the pressures of the gas at the top and bottom of the cylinder; $M$ is the gram molecular weight of the gas (the mass in grams equal to the molecular weight of the gas); $g$ is the constant of gravitation; $R$ is the gas constant; and $T$ is absolute temperature.

[7] The effective weight of a Brownian particle is the excess of its weight over the upward thrust caused by the liquid in which it is suspended.

a number for Brownian particles: any quantity of these particles equal to their molecular weight will contain the same number $N$ of particles. Perrin assumed that this number will be the same as Avogadro's number for molecules.

On the basis of various experiments involving different values for the observable quantities $n'$, $n$, $m$, $h$, and $T$, Perrin used equation (1) to determine a value for Avogadro's number $N$, and discovered that $N$ is indeed a constant, whose approximate value is $6 \times 10^{23}$. He concluded that molecules exist:

> Even if no other information were available as to the molecular magnitudes, *such constant results would justify the very suggestive hypotheses that have guided us* [including that molecules exist], and we should certainly accept as extremely probable the values obtained with such concordance for the masses of the molecules and atoms.... The objective reality of the molecules therefore becomes hard to deny.[8]

Now for antirealist responses. There are two general kinds I shall consider: first, that Perrin's argument for molecules is invalid; second, that even if valid, it does not suffice to establish scientific realism.

## II. THE CIRCULARITY OBJECTION

The first charge is that at the outset Perrin assumes that molecules exist while arguing that they exist. He derives equation (1), which he uses to obtain a value for Avogadro's number $N$ and to see whether $N$ is a constant, from equation (3), which presupposes the existence of molecules. In (3), $n'$, $n$, and $m$ are quantities associated with invisible molecules in a gas. He then assumes that a modified version of (3), namely, (1), can be applied to much larger visible Brownian particles suspended not in a gas but in a dilute fluid.

*Reply.*[9] The argument presented above is only part of Perrin's reasoning, and indeed not the first part. In his book *Atoms*, long before he gets to this argument from Brownian motion, he devotes eighty-two pages to a development of atomic theory, giving chemical arguments in favor of the existence of molecules. His 1909 article begins with a qualitative description of Brownian motion followed by qualitative arguments that Brownian motion is caused by collisions with molecules, and hence that molecules exist. In this article, he writes:

---

[8] *Atoms*, p. 105.

[9] For an expanded discussion of issues in this reply, see my *The Book of Evidence* (New York: Oxford, 2001), chapter 12.

> It was established by the work of M. Gouy (1888), not only that the hypothesis of molecular agitation gave an admissible explanation of the Brownian movement, but that no other cause of the movement could be imagined, which especially increased the significance of the hypothesis.[10]

Guoy performed experiments in which he examined possible external causes of Brownian motion. These included vibrations transmitted to the fluid by passage of heavy vehicles in the street; convection currents produced when thermal equilibrium was not yet attained; and artificial illumination of the fluid. When these and other external sources of motion were reduced or eliminated, the Brownian motion was not altered. Perrin concludes: "these [Brownian] particles simply serve to reveal an internal agitation of the fluid..." (*ibid.*, p. 511).

Perrin offers a second argument from Brownian motion to molecules (*ibid.*, pp. 514-15). Since the Brownian particles are continually accelerating and decelerating, they must be subject to forces exerted upon them, in such a way as to satisfy conservation of momentum. These forces are not present in the particles themselves, nor, as he argued earlier, are they produced by forces external to the fluid. Accordingly, he concludes, they must be produced by the perpetual motions of unobservable particles in the fluid itself.

These arguments for the existence of molecules are presented before Perrin's argument from the law of atmospheres given in section I above. Both involve eliminative-causal reasoning of the following sort:

*Eliminative-causal argument:*

(1) Given what is known, the possible causes of effect $E$ (for example, Brownian motion) are $C$, $C_1,\ldots,C_n$ (for example, the motion of molecules, external vibrations, heat convection currents). (In probabilistic terms, given what is known, the probability is high that $E$ is caused by one of the $C$s cited.)

(2) $C_1,\ldots,C_n$ do not cause $E$ (since $E$ continues when these factors are absent or altered).

So probably

(3) $C$ causes $E$.

Later, I shall consider whether such an argument is valid for molecules (or anything else). At the moment, I am claiming only that an argument of this type is employed by Perrin, and that it is employed prior to, and in addition to, the argument of section I. Perrin offers additional arguments for molecules which are also independent of the law of atmospheres argument.[11]

---

[10] "Brownian Motion and Molecular Reality," pp. 510-11.

Accordingly, independently of his own law-of-atmospheres experiments on Brownian motion, Perrin presents both experimental and theoretical arguments for the following claim:

> $H$: chemical substances are composed of molecules, the number $N$ of which in a gram molecular weight of any substance is approximately $6 \times 10^{23}$.

Let $b$ represent the information, other than the law-of-atmospheres considerations, on the basis of which Perrin infers $H$. He is claiming that the probability of $H$ is high, given $b$, that is,

(4) $p(H/b) > k$

where $k$ is some threshold of high probability, say 1/2.[12]

Now, on the basis of his law-of-atmospheres experiments, using equation (1), Perrin claims that:

> $H'$: The calculation of $N$ done by means of Perrin's experiments on Brownian particles, using equation (1), is $6 \times 10^{23}$, and this number remains constant even when various values in equation (1) are varied.

In $H'$, the number $N$ represents a number for Brownian particles. Perrin can be understood as assuming that, given both $H$ and $b$, the probability of $H'$ is (substantially) increased over what it is given $b$ alone. That is, the probability that Perrin's experiments will yield result $H'$ for Brownian particles is greater given the assumption that molecules exist, and that their number in a gram molecular weight of a substance is constant and approximately equal to $6 \times 10^{23}$, than it is without this assumption. That is,

$p(H'/H\&b) > p(H'/b)$

It follows that

(5) $p(H/H'\&b) > p(H/b)$

---

[11] These involve chemical considerations and also the fact that experiments on various phenomena other than Brownian motion yield approximately $6 \times 10^{23}$ for Avogadro's number, although Perrin claims that for the most part these other methods do not yield the same precision as the determination from Brownian motion. Salmon emphasizes this second part of Perrin's reasoning, which he takes to be an example of "common-cause reasoning." For a discussion of Salmon's particular analysis, see my *The Book of Evidence*, chapter 12.

[12] For arguments that $k$ is greater than or equal to 1/2, see my *The Book of Evidence*.

assuming that neither $p(H)$ nor $p(H'/b)$ is zero. In short, the probability of the molecular hypothesis $H$ is (substantially) increased by $H'$, the results of Perrin's experiments with Brownian motion. From (4) and (5) we can conclude that the molecular hypothesis $H$ is highly probable given $H'$ and $b$, that is,

$$p(H/H'\&b) > k$$

Perrin's argument, so represented, does not involve circularity. Even though Perrin derives the law of atmospheres (3) involving molecules from the assumption that molecules exist, and even though he derives $H'$ from the law of atmospheres for molecules, $H'$ itself does not state or presuppose that molecules exist or that Avogadro's number for molecules is $6\times10^{23}$. $H'$ is established experimentally. On Perrin's view, given $b$, $H'$ bestows substantial probability on $H$ (which does state that molecules exist and that Avogadro's number for molecules is a constant).

## III. THE "MULTIPLICITY" OBJECTION

A second objection is that, given Perrin's experimental results, various hypotheses can be shown to receive just as much support as the particular molecular one that he accepts. And these alternatives need not be committed to the existence of molecules. The multiplicity objection comes in two forms.

A. *Parallel antirealist argument.* The idea, a generalization of one offered by Bas van Fraassen,[13] is that, for each argument used by Perrin whose conclusion is that molecules exist, another argument, at least as good, can be constructed from the same premises which does not conclude that molecules exist, but only that the molecular theory "saves the phenomena." Indeed, an argument of the latter sort is simpler and stronger than the former, since it commits one to much less than does any argument of the former sort. It commits one only to the "empirical adequacy" of the molecular theory, not to its truth, whereas an argument of the former sort is committed to both.

For example, an eliminative-causal argument of the sort attributed to Perrin in section II is replaced by the following:

*Antirealist eliminative-causal argument:*

(1) Given what is known, of the theories that attribute a cause for effect $E$ (for example, Brownian motion), the possible candidates for theories that save the phenomena are ones that invoke causes $C$, $C_1,\ldots,C_n$. (In probabilistic terms, given what is known, the probability is high that at least one of the theories cited saves the phenomena.)

---

[13] *The Scientific Image* (New York: Oxford, 1980).

(2) Theories that invoke causes $C_1,\ldots,C_n$ do not save the phenomena (since these theories predict that, under certain observable conditions, effect $E$ will continue, which new information shows to be false).

So probably,

(3) The theory that invokes cause $C$ (for example, molecules) saves the phenomena.

The conclusion is not that the theory that claims that $C$ causes $E$ is true, or that the entities it postulates (for example, molecules) exist or cause $E$, but simply that it is empirically adequate.

A similar claim is made for any argument that concludes with the truth of a theory postulating the existence of unobservables. For example, with respect to parallel "inference-to-the-best-explanation" arguments, van Fraassen offers two objections (*ibid.*, pp. 20-21). The first is that the claim that scientists reason in accordance with realist versions of such arguments (or analogous eliminative-causal ones) is an *empirical* hypothesis "to be confronted with data, and with rival hypotheses" (*ibid.*, p. 20). van Fraassen proposes a rival hypothesis, namely, that scientists reason in accordance with the antirealist version of inference-to-the-best-explanation (or an analogous version of the eliminative-causal argument given above). Both hypotheses, he seems to be saying, are compatible with the empirical fact that scientists employ explanatory (or eliminative-causal) reasoning. van Fraassen's second objection is that the realist needs an extra premise for his argument, one that van Fraassen regards as false. For inference-to-the-best-explanation, the extra premise is that "every universal regularity in nature needs an explanation" (*ibid.*, p. 21). For causal reasoning, the extra premise would be that every phenomenon in nature has a cause.

*Reply.* (1) To be sure, the fact that Perrin invokes eliminative-causal reasoning does not by itself establish that he is employing a realist or an antirealist version of such reasoning. Here one needs to examine what he in fact says. His claim is not (simply) that of various theories that attribute a cause for Brownian motion, the ones he cites are the possible candidates for saving the phenomena; it is the stronger claim that these theories are the possible candidates for truth. Nor is his conclusion (simply) that the theory invoking molecules as causes in fact saves the phenomena. It is the stronger claim that such a theory is true, that molecules exist ("the objective reality of the molecules therefore becomes hard to deny"), and that molecules cause Brownian motion. If, as van Fraassen insists, it is an empirical question what form of reasoning a scientist in fact is using, then the best empirical method to determine this is to look at what he actually says.

(2) Perrin does not need an extra premise asserting that every phenomenon in nature has a cause (or that every universal regularity needs an explanation). All he needs is the assumption that Brownian motion has a cause. To be sure, he does not explicitly defend this assumption, although he clearly makes it.[14] Presumably, if required to defend this assumption Perrin would have appealed to the idea that Brownian motion involves accelerations of bodies—both changes in speed and direction—which, in the classical physics that he was assuming—require causes. Perrin had no need to introduce some very general assumption that every universal regularity in nature has a cause or needs an explanation (an assumption not even made in classical mechanics). He simply needed to assume that the Brownian motion of the observable particles of gamboge has a cause.

*B. Multiplicity-of-causes objection.* The idea behind this second form of the multiplicity objection is to attack Perrin's particular eliminative-causal argument, which assumes that, given that known observable causes of Brownian motion are eliminated, one can infer an unobservable cause. The problem is that not all possible observable causes have been considered. This objection could be raised as a general criticism of eliminative-causal reasoning ('In general, how do you know you have cited all possible causes of a phenomenon?'). Or it could be raised as a specific one against Perrin ('What reason was there for supposing that Perrin had considered all possible causes of Brownian motion?').

*Reply.* (1) The *general* criticism is based on the assumption that one is justified in employing an eliminative-causal argument only if all but one of the possible causes of the phenomenon have been listed and eliminated. But this is too demanding. One can be justified in employing an eliminative-causal argument if, *given one's background information,* one has considered and eliminated all but one of the possible causes, or at least, all but one of the causes that (on the basis of the background information) have any significant probability of causing the phenomenon in question. The claim that the possible causes cited probably include the actual one can be defended by appeal to the fact that the phenomenon in question is of a certain type that, experience has shown, in other cases is caused by one or the other of the causes cited.

---

[14] See, for example, *Atoms,* pp. 85-86.

(2) The *specific* criticism is that, even if eliminative-causal reasoning can be reasonable, if properly employed, Perrin did not properly employ it. He did not cite all the (plausible) causal candidates for Brownian motion, given information available to him. To be sure, he cites several candidates and argues that experiments eliminate these. But why suppose that these are the only possible or plausible causes permitted by his background information? And even if they are the only possible *observable* causes, why suppose that the motion of particles is the only possible *unobservable* cause?

Here, I suggest, the burden of proof is on the critic. Perrin cites various known causes of motion in a fluid and argues that experiments, particularly those of Christian Wiener in 1863 and Guoy in 1888 and 1889, show that Brownian motion continues unabated when these causes are altered or eliminated.[15] Perrin, quoting Wiener, concludes that the motion "does not originate either in the particles themselves or in any cause external to the liquid, but must be attributed to internal movements...."[16] Since Perrin cited and eliminated various possible causes, it is, I think, up to the critic to say what other possible causes he should have eliminated, given his information (and why these should be assumed to be observable).

(3) Both the general and the specific criticisms cited above, if valid, could equally well be used against the antirealist version of the eliminative-causal argument. That version cites theories that invoke causes $C$, $C_1$,...,$C_n$ as the possible candidates for saving the phenomena and then eliminates all but one of these theories. But, the critic can ask, why suppose that these are the only possible causal theories that will save the phenomena? And the critic can ask Perrin the specific form of this question for Brownian motion. So an antirealist version of eliminative-causal reasoning is no better off than the realist version. My response is simply to reject the critic's claims for the reasons given in replies 1 and 2 above.

---

[15] Wiener demonstrated that Brownian motion is not caused by infusoria (one-celled animals found in exposed bodies of water), by electrical forces, by temperature differences, or by evaporation of the fluid. In addition to these, Guoy's experiments eliminated causes pertaining to the size, composition, and density of the Brownian particles. For a discussion of various observable causes of Brownian motion postulated by scientists opposed to molecular theory, see Nye, *Molecular Reality* (London: Macdonald, 1972), pp. 22-27. Nye concludes (p. 27) that in his experiments, "Guoy [in 1889] refuted, point by point, all previous theories of Brownian movement other than the [molecular-] kinetic."

[16] *Atoms*, p. 86.

## IV. THE "LIMITS OF EXPERIENCE" OBJECTION

The general idea here is simple and powerful. By observing nature, we can only make inferences about what is observable in nature, not what is unobservable. Here are claims of two prominent antirealists. First, Pierre Duhem[17]:

> Now these two questions—Does there exist a material reality distinct from sensible appearances? and What is the nature of this reality?—do not have their source in experimental method, which is acquainted only with sensible appearances and can discover nothing beyond them (*ibid.*, p. 10).

Second, van Fraassen[18]:

> I explicate the general limits [of experience] as follows: *experience can give us information only about what is both observable and actual* (*ibid.*, p. 253).

The claim, then, is that any argument from what is observed to the truth of claims about what is unobservable is unjustified. Since Perrin claimed truth for his theory about unobservable molecules, and since his argument is based on observed results of experiments that he and others conducted, his argument is unjustified.

What is the basis for the claim that by observing nature one can make inferences only about what is observable? A simple defense, one indeed suggested by Duhem, is that this is the essence of empiricism; inferring beyond the observable is metaphysics, which is not empirical. Indeed, immediately following the passage quoted above, Duhem claims that the resolution of questions as to what lies beyond the "sensible appearances" "transcends the methods used by physics; it is the object of metaphysics" (*op. cit.*, p. 10). But this begs the issue, which in this case is that of characterizing empiricism. Perrin, no less than Duhem, regarded himself as engaged in empirical science, not metaphysics. He believed that empirical arguments could be given for the truth of a theory about unobservable molecules. It will not do simply to claim that his argument is not (entirely) empirical, or that he has gone beyond what empiricism allows.

van Fraassen regards as extreme a policy that permits inferences beyond the "range of possible additional evidence," that is, inferences to the truth of claims about unobservables.[19] If only evidence can justify a belief, he goes on to say, then belief in the truth of a theory is "supererogatory," since we can have evidence for truth only via evidence for empirical adequacy (*ibid.*, p. 255). van Fraassen's first claim presupposes, or at least strongly suggests, that inferences to the truth of claims about unobservables are inferences beyond what the evidence allows. His second claim is that one can have evidence only for the empirical adequacy, not the truth, of a theory postulating unobservables, or that evidence for truth only amounts to evidence for empirical adequacy. Both claims, I suggest, require justification. Perrin, for example, clearly believed that he had evidence for the truth of propositions about unobservable molecules, not simply for their empirical adequacy; he had evidence that molecules exist, not simply evidence that the molecular theory saves the phenomena. Why was he wrong in this belief?

---

[17] *The Aim and Structure of Physical Theory* (Princeton: University Press, 1991).

[18] *Images of Science*, Paul M. Churchland and Clifford A. Hooker, eds. (Chicago: University Press, 1985).

[19] *Images of Science*, p. 254.

Let me suggest an argument an antirealist might offer in the spirit of the two passages quoted above. Suppose that having observed a great many *A*s and found them all to be *B*, one infers that all *A*s are *B*. For example, from the fact that all observed bodies have mass one infers that all bodies, including any unobservable ones, do too.[20] (In Perrin's case, we might consider an inference from 'All observed accelerating bodies in contact with other bodies exert forces on them' to 'All accelerating bodies, including molecules (if any exist), in contact with other bodies exert forces on them'.) Is an inference of this sort legitimate? Not necessarily. The sample observed may be unrepresentative. Even if we do not know that it is unrepresentative, we may have no positive empirical reason to think it is not biased. Suppose that all the *A*s chosen for observation satisfy a condition *C*. Then unless we can argue that *C* is irrelevant—that *C* does not bias the sample with respect to *B*—we should infer not that all *A*s are *B*, but, at best, that all *A*s that satisfy *C* are *B*.

---

[20] The points to be made below will apply as well to causal reasoning from all observed *A*s (for example, accelerations) are caused by *B*s (for example, forces) to all *A*s are caused by *B*s Realists generally extend inductive generalizations and causal reasoning to unobservables. For example, Isaac Newton writes, in his discussion of his inductive Rule 3: "The extension of bodies is known to us only through our senses, and yet there are bodies beyond the range of these senses; but because extension is found in all sensible bodies, it is ascribed to all bodies universally"—*Principia*, I.B. Cohen and Anne Whitman, eds.(Berkeley: California UP, 1999), Book 3, p. 795. Other such properties Newton mentions in this discussion are hardness, impenetrability, mobility, inertia, and gravitational attraction. Newton also extends his causal Rule 2 to apply to reasoning from effects to unobservable causes. Antirealists such as Duhem and van Fraassen do not deny the existence of unobservables. Their claim, contrary to Newton's, is that one cannot extend inductive or causal reasoning to unobservables by claiming truth for the conclusions of such arguments.

Now, says the antirealist, let *C* be the condition of *being observable*. All observed *As* satisfy this condition. Indeed, one can never observe an *A* that fails to satisfy it. So, unless one can demonstrate that this does not bias the sample of *As* observed with respect to *B*, from the fact that all *As* observed have been *B*, all we can legitimately infer is that all observable *As* are *B*. The only way of demonstrating that "observability" does not bias the sample of *As* observed with respect to *B* is by collecting a suitable sample of unobservable *As* and showing that in this sample the unobservable *As* are all *B* (or perhaps by showing that in other cases, samples of unobservables match observables with respect to some property). But, of course, one cannot do that! One cannot observe unobservables. So in order to exclude potential bias, we are restricted to making (inductive and causal) inferences to what is observable. This argument can be extended to cover all types of nondemonstrative inferences. That is why Duhem is right in claiming that we can discover nothing beyond appearances, and why van Fraassen is justified in claiming that experience can give us information only about what is observable.

*Reply*. An antirealist who argues in the previous manner looks at the situation as one involving so-called *stratified sampling*. The population of *As* is divided into two classes or strata: the observables and the unobservables (if any). To make inferences about the entire class of *As* with respect to a property *B*, the antirealist is claiming, one needs to select randomly members from both strata for observation. Since one cannot select unobservable members for observation, one cannot legitimately, without potential bias, make inferences about this stratum, but only about the observable stratum.

If this argument is derived from the general principle that to make an inductive inference about a class, samples must be taken from all strata of that class, then no inference from any observed sample will be possible. Let the stratifying condition *C* be 'has been observed' (or 'has been observed prior to 2005'). All the *As* observed satisfy condition *C*. We cannot (now) take a sample from the subclass satisfying not-*C*. Accordingly, if the general principle of stratification for any condition *C* is correct, then no generalizations are possible.

Assuming that the antirealist does not accept this general principle with respect to all stratifying conditions *C*, what reason can he offer for supposing that the specific stratifying condition 'being observable' is a biasing condition? Does he have any *empirical* reason for supposing that, in general, when considering a class of *As*, the unobservable stratum of *As*, if any exist, is different with respect to *B* from the observable stratum? No, he does not, since, by his own admission, unobservables cannot be sampled. Nor does he have some a priori argument showing that the two strata are different with respect to *B*. All he has is the weak a priori claim that the strata *may* be different—a claim about a mere logical possibility. But this is not sufficient to justify the methodological injunction that to make a legitimate inductive inference about the entire class consisting of observables and unobservables one must sample both strata. Indeed, if it were sufficient, the antirealist would be prevented from making any generalizations from what has been observed to what is observable but never observed. If one stratifies the class of observable *A*s into those which are or will be observed and those which never will be observed, then it is logically possible that these strata are different with respect to *B*.

The realist, who wants to make inferences about unobservables as well as observables, must also reject the general principle of stratification above. Can he do so without rejecting the idea that sampling requires variation? Can the realist offer an argument that provides support for his claim that "observability" is not, in general, a biasing condition?

I suggest that here the realist has an advantage over the antirealist. The realist can offer an empirical argument that provides support for his claim. There is a type of variation that can be produced for the purpose of meeting the antirealist challenge. One can vary conditions or properties in virtue of which something is observable (or unobservable). For example, items can be observable (or unobservable) in virtue of their size, their distance from us in space or time, their duration, their interactions (or lack of them) with other items, and so on. Suppose that all physical bodies that have been observed have mass. Have we biased the observed sample with respect to mass by observing only bodies that are observable? Observed physical bodies are necessarily observable. Therefore, we cannot vary the physical bodies observed by observing unobservable as well as observable ones. But we can vary properties of physical bodies in virtue of which they are observable (or unobservable).

Suppose we do the latter and find that bodies have mass even when we observe bodies that have different sizes, different distances from us, different durations, and different numbers and kinds of interactions with other bodies. In the absence of any contrary empirical information, we can then infer that size, distance, duration, and numbers and kinds of interactions do not alter the situation: bodies observed with diferent sizes, distances, durations, and interactions all have mass. So we infer that differences in these properties—differences that make some bodies observable and others not—make no difference as far as having mass is concerned.[21] If we vary the conditions in virtue of which bodies are observable and find no differences in whether bodies have mass, and if we have no contrary empirical information, then we have offered an empirical argument to support the claim that the fact that all observed bodies are observable does not bias the observed sample with respect to the property of having mass. In so doing, we have provided empirical grounds for inferring that all bodies have mass, whatever their size, distance, duration, and so on, and hence, whether or not they are observable.[22]

An antirealist may vehemently repudiate this argument by insisting that the only way to vary a condition *C* that is satisfied by all observed *A*s is to observe *A*s that satisfy *C* and *A*s that do not. The antirealist will then reject the claim that the types of variations above (which do not do this) suffice. But the claim that this is the only type of variation that can eliminate bias needs an argument, especially since the antirealist should allow variations in size, distance, and so on to count as varying the conditions when making inferences about observables.

For example, suppose an antirealist makes an inference he considers legitimate from the fact that all observed bodies have mass to all *observable* bodies have mass. In doing so, in order to preclude bias, one might vary the size of the bodies examined, their distance from other bodies, and a host of other properties. More generally, some variations involve changing "qualitative" properties by examining *A*s that have such properties and *A*s that do not. But there are also variations that involve changing "quantitative" properties by examining *A*s with varying amounts of such properties. If the antirealist rejects this second type of variation for eliminating bias, he needs an argument for doing so. If he accepts it, but only for inferences to observables, again he needs an argument for this restriction. The claim that unobservable bodies *may* be different from the observed ones with respect to having mass—in the sense of logical possibility—cuts no ice, since exactly the same could be said for the observable (but not yet observed) bodies. Nor, again, does the antirealist have either an empirical or an a priori argument to show that the observed bodies are unrepresentative of the unobservable ones with respect to having mass.[23]

---

[21] This conclusion is subject to the restriction noted above, namely, "in the absence of any contrary information." With mass, there is no such contrary information. But with many properties, there are; for example, observed bodies of various sizes, distances from us, durations, and interactions have temperature. But we have other empirical information, including an empirically supported theory about the nature of heat, that prevents the inference that bodies of all sizes, and so on have temperature (for example, individual molecules). In such cases, however, it is not the unobservability of such bodies that prevents the inference that all bodies have temperatures, but information about the nature of heat. More generally, with the addition of quantum-mechanical information, unknown of course to Perrin in 1908, various conclusions from observations made on observed macrobodies cannot be drawn to bodies of much smaller size such as electrons. But this is not because such bodies are unobservable but because bodies of that size exhibit quantum-mechanical properties.

[22] This type of reasoning is analogous to what Philip Kitcher, in a recent perceptive article on realism, calls "the Galilean strategy"—"Real Realism: The Galilean Strategy," *The Philosophical Review*, cx (April 2001): 151-97, particularly pp. 173-80.

## V. IS THIS SCIENTIFIC REALISM?

The final objection, possibly the most important, is that whatever Perrin proved in his argument, even if by means of his experiments he did prove that molecules exist, this is not scientific realism. An objection of this sort will be raised by anyone, realist or antirealist, with a more demanding view of what scientific realism requires. Various candidates for scientific realism have been suggested in recent years by proponents as well as critics; three prominent ones will be noted.

(1) Scientific realism is a view about truth and reference in scientific theories generally. For example, Richard Boyd[24] and Stathis Psillos,[25] defenders of scientific realism, and Larry Laudan,[26] a severe critic, all claim that scientific realism involves a number of central theses, including these:

(A) Scientific theories (at least in the "mature" sciences) are typically approximately true, and more recent theories are closer to the truth than other theories in the same domain.

(B) The observational and theoretical terms within the theories of a mature science genuinely refer (roughly, there are substances in the world which correspond to the ontologies presumed by our best theories).

(C) Successive theories in any mature science will be such that they "preserve" the theoretical relations and the apparent referents of earlier theories (that is, earlier theories will be "limiting cases" of later theories).

(D) Acceptable new theories do and should explain why their predecessors were successful insofar as they were successful.

---

[23] To be sure, there are variations that can be made to show that mass is not a constant but varies with velocity. But this fact about mass has nothing to do with the observability or unobservability of the bodies with mass.

[24] "The Current Status of Scientific Realism," in Jarrett Leplin, ed., *Scientific Realism* (Berkeley: California UP, 1984), pp. 41-82.

[25] *Scientific Realism* (New York: Routledge, 1999), p. xix.

[26] "A Confutation of Convergent Realism," in Leplin, pp. 218-49.

This is Laudan's formulation (*op. cit.*, pp. 219-20). To these theses Boyd adds:

> (E) The reality which scientific theories describe is largely independent of our thoughts or theoretical commitments (*op. cit.*, p. 42).

Thesis (E) relates scientific realism to a core idea of what is sometimes called *metaphysical realism*, namely, that there is a mind- and theory-independent way the world is; to this core idea it adds the claim that scientific theories describe such an independent world.[27]

Even if Perrin proved that molecules exist, he did not prove theses (A)-(E), or even make them probable. Nothing Perrin did establishes or makes probable claims about scientific theories generally, or even about ones in mature sciences. Nor did Perrin establish that there is a mind- and theory-independent world or that theories in the mature sciences describe such a world approximately correctly.

(2) Scientific realism is a view about the *aim* of science. Here is van Fraassen's formulation:

> *Science aims to give us, in its theories, a literally true story of what the world is like; and acceptance of a scientific theory involves the belief that it is true.* This is the correct statement of scientific realism.[28]

van Fraassen distinguishes the aim of science, as a type of activity, from the aims of particular scientists, which may include fame and fortune.[29] Now, proving that molecules exist does not establish that the aim of science is to provide a literally true story of the world (or to satisfy any other broad conditions, such as describing natural kinds and dependencies), not even if this were Perrin's aim in the case of molecules. One would need to show more than this to establish scientific realism in the present sense. (I shall return to this in section VII.)

---

[27] Psillos divides scientific realism into a "metaphysical" part that "asserts that the world has a definite and mind-independent natural-kind classification"; a "semantical" part that says that scientific theories are capable of truth values and that theoretical terms are capable of denoting unobservable entities in the world; and an "epistemic" part that claims that "mature and predictively successful scientific theories [are] well-confirmed and approximately true of the world" and that the theoretical terms they employ denote entities that exist.

[28] *The Scientific Image*, p. 8. van Fraassen's contrasting antirealism is this: "Science aims to give us theories which are empirically adequate, and acceptance of a theory involves as belief only that it is empirically adequate." Note that such an antirealism is compatible with the truth of theses (A),(B),(C), and (E) of the Boyd-Psillos-Laudan formulation of realism, though, van Fraassen would emphasize, it is not the aim of science to produce theories that satisfy these theses but only to produce theories that are empirically adequate.

[29] Some realists insist on broader aims than the one van Fraassen assigns to realists. For example, Kitcher defends an account of realism that combines certain elements of van Fraassen's definition with that of the Boyd-Psillos-Laudan account—*The Advancement of Science* (New York: Oxford, 1993), chapter 5. Kitcher's realism, as expressed in this book, includes the idea that the aim of science is to recognize natural kinds and furnish a set of explanatory schemata that pick out dependencies in the world; it also includes the claim that various parts of the sciences achieve these aims. For his more recent account, see "Real Realism: The Galilean Strategy."

(3) The scientific realism of interest to philosophers is not itself an *internal* scientific question, to be settled by scientific reasoning, but an *external* one concerning the adequacy of the scientific representation of the world. It cannot be established by empirical means. Both Rudolf Carnap[30] and Arthur Fine[31] have defended a distinction between internal and external questions. Their views about internal questions are somewhat similar, although they take very different positions on external questions.

For Carnap, the question of whether molecules exist can be approached in two ways. First, it can be treated as an empirical question within what Carnap calls a "linguistic framework," which contains rules of language and of inference, including rules governing what counts as evidence for what. Considered as an internal question within a framework permitting "theoretical terms" for unobservable entities, the answer to whether molecules exist can be determined empirically by the sorts of arguments from experiments Perrin provided. But this is not what philosophers usually have in mind when they assert (or deny) scientific realism with respect to theoretical entities such as molecules. For these philosophers, the question is an external one concerning the adequacy of the "theoretical entity" framework within which molecular claims are made. For Carnap, the claim that the framework is adequate is not a claim about its relationship to a framework-independent world. It is a pragmatic claim about the employability of the framework based on features such as simplicity, ease of use, and familiarity. Accordingly, different frameworks can be adopted, some of which have no terms for unobservables such as molecules. Some frameworks may be more user-friendly than others, but none is "correct" or "incorrect."

Carnap's internal questions are part of what Fine calls the *natural ontological attitude* (NOA). So, Fine can agree, Perrin established the existence of molecules by means of his experiments. In this internal sense, an antirealist can agree with a realist about what Perrin accomplished scientifically. What a realist does, according to Fine, is to step outside of scientific activity and claim that theories correspond to reality. Fine does not endorse Carnap's line that such a claim is pragmatic. Rather, he says simply that what the realist is doing when he steps outside of scientific activity is tantamount to pounding the table and saying 'Molecules exist, really!'. This adds only emphasis to the internal claim. Fine, then, advocates NOA, which he regards as different from both realism and antirealism but as something that realists as well as antirealists can accept.

---

[30] "Empiricism, Semantics, and Ontology," in Carnap, *Meaning and Necessity* (Chicago: University Press, 1956), pp. 205-21.

[31] "The Natural Ontological Attitude," in Leplin, pp. 83-107.

Both Carnap and Fine could agree that Perrin proved empirically that molecules exist, where the latter claim is understood as one internal to science (as part of NOA). For Carnap and Fine, Perrin did not prove, nor could he, that a framework containing terms for molecules corresponds to reality in some external sense. So if, as is typical, scientific realism is construed as an external doctrine, Perrin did not establish scientific realism. For Fine, neither Perrin nor anyone else could establish realism in an external sense. For Carnap, Perrin showed, by using the framework he did, at best that it was useful; he did not show that it is the only or the most useful one for dealing with questions involving Brownian motion.

These three views about scientific realism represent a spectrum of positions. Perrin's arguments do not establish scientific realism in a sense of that term advocated by any of the philosophers noted. So is Salmon mistaken in his claim that Perrin provided a valid experimental argument for scientific realism? To answer, some history is relevant.

## VI. HISTORICAL REASONS FOR REJECTING ATOMISM

We need to consider reasons that certain scientists rejected atomic-molecular theory until the first decade of the twentieth century. These reasons were known to Perrin, who responded to them. Critics of atomic theory included Duhem, Ernst Mach, Friedrich Wilhelm Ostwald, Henri Poincaré, and Max Planck. Grounds cited for rejecting the theory were in part scientific.[32] But they also included important philosophical or methodological reasons. One was the claim that physicists such as James C. Maxwell and Ludwig Boltzmann, who proposed atomic-molecular theories, were employing an illegitimate "method of hypothesis." From unproved hypotheses about atoms and molecules, they were deducing observable consequences and claiming that from the truth of the latter, one could infer the truth of the hypotheses themselves.[33] While Mach, for example, did not repudiate the use of hypotheses, he did reject the idea that one could infer their truth or probability from true observed consequences. He regarded the hypotheses of the atomic theory as provisional "mental artifices" for summarizing known observational facts and facilitating observational predictions.[34] Once such hypotheses have served their organizational and predictive purposes, they are to be discarded, not accepted as true or probable. Indeed, Mach championed an antirealist "sensationalist" view about observable matter, which he regarded as complexes of sensations.[35]

---

[32] For example, atomic theories, as proposed, were purely mechanical theories that should entail reversible processes; but the latter are incompatible with observed thermodynamic phenomena.

A related philosophical claim made by some of the critics of atomic theory (whether or not they were Machian reductionists) was that hypotheses about unobservables can never be established as true or probable by empirical means; such hypotheses can never be known to be true. For example, Poincaré[36] speaks of the atomic hypothesis as "indifferent," meaning that although it may be useful, it can never be empirically established or shown to be more probable than rival hypotheses that assert the continuity of matter (*ibid.*, pp. 152-53). For Duhem,[37] as noted earlier, science can know only "sensible appearances and can discover nothing beyond them" (*ibid.*, p. 304). Unlike Poincaré and Mach, Duhem does not even regard atomic hypotheses as useful devices for summarizing and predicting observable phenomena.

Like some of the critics of the atomic theory, and no doubt because of their criticism, Perrin offers his own general philosophical/methodological reflections on how to proceed with scientific investigations. He distingishes two scientific methods, which he calls the *inductive method* and the *intuitive method*.[38] The former, associated with the critics of atomic theory noted above, begins with what is observed and reasons only to statements about "objects that can be observed and to experiments that can be performed" (*ibid.*, p. vii). The second method infers the existence of an unobservable mechanism:

---

[33] Whether Maxwell and Boltzmann were in fact doing this is questionable. For a discussion of Maxwell's reasoning, see my *Particles and Waves*.

[34] He writes: "However well fitted atomic theories may be to reproduce certain groups of facts, the physical inquirer who has laid to heart Newton's rules will only admit those theories as *provisional* helps, and will strive to attain, in some more natural way, a satisfactory substitute. The atomic theory plays a part in physics similar to that of certain auxiliary concepts in mathematics; it is a mathematical *model* for facilitating the mental reproduction of facts"—*The Science of Mechanics* (LaSalle, IL: Open Court, 1960), p. 589.

[35] "...ordinary 'matter' must be regarded merely as a highly natural, unconsciously constructed mental symbol for a complex of sensuous elements..."—*Contributions to the Analysis of the Sensations* (LaSalle, IL: Open Court, 1890), p. 152.

[36] *Science and Hypothesis* (New York: Dover, 1952).

[37] "Physics of a Believer," in *Aim and Structure of Physical Theory*.

[38] *Atoms*, p. vii.

> In studying a machine, we do not confine ourselves only to the consideration of its visible parts, which have objective reality for us only as far as we can dismount the machine. We certainly observe these visible pieces as clearly as we can, but at the same time we seek to divine the *hidden* gears and parts that explain its apparent motions.
>
> To divine in this way the existence and properties of objects that still lie outide our ken, *to explain the complications of the visible in terms of invisible simplicity*, is the function of intuitive intelligence which, thanks to men such as Dalton and Boltzmann, has given us the doctrine of Atoms. This book aims at giving an exposition of that doctrine (*ibid.*, p. vii).

Perrin claims that in the times in which he is writing, the method of intuition has gone ahead of induction in rejuvenating the doctrine of energy by the incorporation of statistical results borrowed from atomists (*ibid.*, p. viii).

In these introductory passages, Perrin does not describe the intuitive method in detail other than to say that it is a method for inferring an invisible underlying reality from visible things and events in such a way that the former explains the latter. I suggest that we take his own arguments for molecules from experiments on Brownian motion as representing his use of the "intuitive method." If so, then that method is not simply a matter of speculating about an invisible realm. Nor is it a Machian provisional "mental artifice" for representing the facts, to be discarded once it has served its purpose. Nor is it a method simply for inferring the truth of claims about what can be observed (that is what he calls the inductive method). Nor, by contrast with Mach, does Perrin say or imply that assertions about ordinary matter are to be understood as claims about sensations; or (with later logical positivists) that assertions about an invisible realm of molecules are to be understood as claims about a visible realm such as that of Brownian motion; or that when he infers that molecules are real and that certain claims about them are true, by 'real' and 'true' he means 'useful' or 'saves the phenomena'.

Finally, Perrin's intuitive method is not the method of hypothesis, which infers the truth or probability of an hypothesis simply from the fact that the hypothesis explains and predicts observable phenomena. To be sure, Perrin's full argument involves citing explanatory chemical reasons from combinations of elements and compounds. But in addition, it involves eliminative-causal reasoning of a sort indicated in section II above. It involves an appeal to similar determinations of

Avogadro's number *N* from a variety of experiments on phenomena other than Brownian motion. And, perhaps most importantly, it involves a calculation of *N* from Perrin's own experiments on Brownian motion and a demonstration that this number remains constant (proposition *H'* in section II). The former arguments, according to Perrin, show that the molecular hypothesis is highly probable. The latter sustains and even increases this probability on the basis of new, precise, experimental results.

Now we are in a position to respond to Salmon's challenge.

## VII. DID PERRIN PROVIDE AN EXPERIMENTAL ARGUMENT FOR SCIENTIFIC REALISM?

Or did he merely furnish an answer to what Carnap calls an internal question? That is, is Perrin's argument more than simply an empirical argument for the existence of a particular natural kind, namely, molecules—entities which had been previously postulated by scientists but not shown to exist? It is this and a good deal more. It is an argument to the existence of something that was regarded as dubious or objectionable on philosophical or methodological grounds. These grounds included the idea that from what is observed, one can make valid inferences only to what is observable, not to what is unobservable. Moreover, they are grounds for rejecting unobservables generally, not just molecules. (Duhem, for example, rejected a range of physical theories postulating unobservables.)

Perrin did not respond to these critics by making molecules observable. Rather, he claimed that there is a reasonable method or mode of reasoning (the intuitive method) that starts with observed experimental results and can be used to infer the existence of things that are unobservable (or 'invisible', or 'hidden', to use Perrin's terms). He showed in detail how this method could be used to infer the existence of invisible molecules from experimental observations on Brownian motion. And he clearly regarded this intuitive method as a general one, not restricted simply to molecules.[39] This is not just scientific business as usual. It is not simply an argument establishing (or rejecting) the existence of just one more type of physical entity. Nor did Perrin or his antirealist opponents take this as scientific business as usual.

The scientific realism implicit in Perrin's arguments can be put like this:

---

[39] "The use of the intuitive method has not, of course, been used only in the study of atoms, any more than the inductive method has found its sole application in energetics"—*Atoms*, p. vii.

(1) There are unobservables (for example, molecules).

(2) Their existence and their properties can be inferred (only) on empirical grounds, in some cases from experiments, so that a claim to know they exist and have these properties is justified.[40]

(3) A legitimate mode of reasoning that can be used for this purpose involves two important components:
   (a) causal-eliminative reasoning to the existence of the postulated entity, and to certain claims about its properties, from other experimental results;
   (b) an argument to the conclusion that the particular experimental results obtained are very probable given the existence of the postulated entity and properties.

Part (a) in (3) is intended to establish the high probability for the claim that the entity exists and has certain properties. Part (b) can be shown to sustain and possibly increase that probability on the basis of the new experiments. Reasoning involving (a) and (b) can be used to infer the existence of entities whether observable or unobservable.

The scientific realism reflected in points (1)-(3) does not say that theories about unobservables in mature sciences are (approximately) true and that the unobservables postulated exist, or that the aim of science is to obtain true theories about unobservables as well as observables. Nor does it adopt either some metaphysical viewpoint external to science, saying that there is a mind- and theory-independent way the world is and/or emphatically endorsing the adequacy of a scientific representation of this world, or a pragmatic viewpoint stressing the usefulness of scientific theorizing. It is not scientific realism *in these senses*.

Nor is it a restricted internal realism that says simply that molecules (electrons, or whatever) exist. It is much more general than this, and it has a methodological as well as an ontological component. It claims that there are unobservables (Perrin's realm of "invisible simplicity"), and that valid arguments can be used to infer their existence and properties; these are entities which, on philosophical and/or methodological grounds, antirealists reject or understand in some nonrealistic way. It says, however, the arguments for these entities and claims about their properties are empirical ones, so that what entities exist in this realm and what properties they have are empirical questions. The particular empirical argument for a given unobservable will depend on, and vary with, the unobservable postulated. No *general* empirical argument can be given for all unobservables postulated. Nor can the issue of the existence of unobservables be settled a priori. Nevertheless, there is an a priori assumption that is essential to scientific realism so understood, namely, that a valid mode of reasoning can be employed (such as (a) and (b) in point (3) above) which can justify a belief in the truth of propositions about the unobservable entities inferred.

---

[40] Statement (2) is close to what Leplin calls "minimal epistemic realism," a doctrine he defends in *A Novel Defense of Scientific Realism* (New York: Oxford, 1997).

Does this deserve the name "scientific realism"? Fine may answer "no," claiming that it does not go beyond what his neutral NOA permits. Fine says that NOA, and hence realism and antirealism, "accept(s) the results of scientific investigations as 'true' on par with more homely truths."[41] Hence, he will insist, since the results of scientific investigation include empirically based inferences to the existence of unobservables such as molecules, NOA, together with realism and antirealism, accepts both (1) and (2). Moreover, he may add that NOA, and hence both realism and antirealism, accepts point (3), since, he says, NOA sanctions "ordinary relations of confirmation and evidential support, subject to the usual scientific canons" (*ibid.*, p. 98) (although he does not formulate any such canons, or indicate whether they include the type of inference in point (3)). But if NOA does endorse points (1)-(3), then, I suggest, *it endorses a form of scientific realism rejected by scientists of the sort Perrin was opposing who are generally classified as antirealists.*[42] These scientists, including Duhem, Mach, Ostwald, and Poincaré, claim either that point (1) is false or else that, if it is true, it is unknowable by empirical means. (They do not regard propositions about unobservables on a "par with more homely truths.") And they reject points (2) and (3), since they raise general methodological objections to inferences from what is observed to what is unobservable.

Moreover, if an antirealist says either that unobservables do not exist or that, if they do, they are empirically unknowable, and if he is an empiricist about science, then, as in the case of Duhem and van Fraassen, he is likely to deny that the aim of science is to provide true theories about unobservables. This is because he holds that such an aim cannot be satisfied at all or cannot be satisfied by empirical means. Although this position does not follow deductively from a denial of points (1)-(3), it is a natural claim to make if one wants to retain empiricism in science and hold that the aim of science can be achieved. Similarly, if one is a realist and states that unobservables do exist and that claims to know this in particular cases can be empirically justified, then it is natural to assert that the aim of science, or at least one of its aims, is to provide true theories about such unobservables.

---

[41] Fine, "The Natural Ontological Attitude," p. 96.

[42] Both Alan Musgrave—"NOA's Ark: Fine for Realism," *Philosophical Quarterly*, xxxix (1989): 383-98—and Psillos—*Scientific Realism*, chapter 10—make the general argument that Fine's NOA is not neutral, but a realist position that is incompatible with certain standard antirealist views.

The most important reason that antirealists such as Duhem and van Fraassen have for saying that the aim of science is (a) to provide theories that save the phenomena, rather than (b) to provide theories that are true, is that they regard (a) but not (b) as doable and empirically justifiable. By contrast, realists such as Perrin and Newton regard both (a) and (b) as doable and empirically justifiable. Since they do, and since they also regard unobservables (such as molecules and universal gravity, which Newton considered a force extended to bodies "beyond the range of the senses") as causally responsible for observable phenomena (such as Brownian motion and motions of the planets), and since they regard particular causal claims of the latter sort as justifiable empirically, it would be natural for them to hold that:

(4) One of the aims of science is to provide (approximately) true theories of what the world is like.

where truth applies to unobservables as well as observables. Point (4) is indeed strongly suggested by the methodological remarks of Newton (for example, in his Rules 1 and 4)[43] and Perrin (in his intuitive method). Accordingly, since it is not entailed by realist points (1)-(3), it is reasonable to consider it part of the realist position. Its truth is not demonstrated by Perrin's empirical argument for molecules, but it is plausible to say that Perrin presupposed its truth in conducting his investigation into the cause of Brownian motion.

The same cannot be said, however, for the doctrine of realism defended by Boyd and Psillos and attacked by Laudan. Points (1)-(4) make no claims about whether, in general, scientific theories in the mature sciences are true, contain terms that refer to objects that exist, and describe a mind- and theory-independent world; such a realism is much stronger than (1)-(4). Nor were any such general claims about theories in the mature sciences made or presupposed by Perrin in his investigation (or by Newton in his argument for the existence of a universal gravitational force).

---

[43] Rule 1 requires that causes postulated be "both true and sufficient to explain their phenomena"; Rule 4 requires that "propositions gathered from phenomena by induction should be considered either exactly or very nearly true...until yet other phenomena make such propositions either more exact or liable to exceptions." Newton applies both rules to unobservables as well as observables.

For scientific realism *of the sort that was supported by scientists such as Perrin and rejected by his opponents* points (1)-(4) suffice. These scientists claimed (or denied) that there is an empirically knowable realm of unobservables responsible for observable effects, and since there is, one of the aims of science is to provide true theories about this realm. Whether any particular theory is true is to be determined by empirical considerations specific to that theory. From the fact that the theory is part of a mature science one cannot infer that it is true or probable. Indeed, as Laudan has argued, historically many such theories have been empirically refuted. For scientific realism of the kind under attack by Duhem, Mach, and other scientists (as well as by van Fraassen), and defended by Perrin (and by Salmon), what is important is the idea that there is a realm of unobservables, claims about which can be empirically justified as true. Whether unobservables do exist, and if so which ones, and what properties they have, are issues to be determined by empirical arguments of the sort Perrin provided for molecules. Accordingly, I regard Salmon's conclusion as justified. In an historically and conceptually important sense of "scientific realism" (though not in every sense assigned to that term), Perrin's experimental argument for molecules provides an empirical basis for scientific realism.

PETER ACHINSTEIN

Johns Hopkins University

## NEW BOOKS: PAPERBACK REPRINTS

ARRINGTON, ROBERT L., ed.: *A Companion to the Philosophers.* Malden, MA: Blackwell, 1999/2001. xix, 696 p. Paper $19.99.

BARRETT, JEFFERY A.: *The Quantum Mechanics of Minds and Worlds.* New York: Oxford, 1999/2001. xv, 267 p. Paper $21.95.

BENJAMIN, WALTER: *The Arcades Project.* HOWARD EILAND and KEVIN MCLAUGHLIN, trans. Cambridge: Harvard, 1999/2002. xiv, 1073 p. Paper 22.95.

BERLIN, ISAIAH: *Against the Current: Essays in the History of Ideas.* HENRY HARDY, ed. Princeton: University Press, 1979/2001. liii, 419 p. Paper $19.95; *The Power of Ideas.* HENRY HARDY, ed. Princeton: University Press, 2000/2002. xv, 240 p. Paper $16.95.

BOBZIEN, SUSANNE: *Determinism and Freedom in Stoic Philosophy.* New York: Oxford, 1998/2001. xi, 441 p. Paper $26.95.

BRANDOM, ROBERT B.: *Articulating Reasons: An Introduction to Inferentialism.* Cambridge: Harvard, 2000/2001. 230 p. Paper $18.95.

CHIHARA, CHARLES S.: *The Worlds of Possibility: Modal Realism and the Semantics of Modal Logic.* New York: Oxford, 1998/2001. xii, 342 p. Paper $26.00.

CURTIN, DEANE: *Chinnagounder's Challenge: The Question of Ecological Citizenship.* Bloomington: Indiana UP, 1999/2001. xviii, 218 p. Cloth $24.95, paper $14.95.

DAVIDSON, DONALD: *Inquiries into Truth and Interpretation.* New York: Oxford, 1984/2001. xxiii, 296 p. Paper $21.95.

DELEUZE, GILLES: *Empiricism and Subjectivity: An Essay on Hume's Theory of Human Nature.* CONSTANTIN V. BOUNDAS, trans. New York: Columbia, 1991/2001. xi, 163 p. Paper $17.50.

DWORKIN, RONALD: *Sovereign Virtue: The Theory and Practice of Equality.* Cambridge: Harvard, 2000/2002. 511 p. Paper $19.95.

FEYERABEND, PAUL: *Conquest of Abundance: A Tale of Abstraction versus the Richness of Being.* Chicago: University Press, 1999/2001. xviii, 285 p. Paper $15.00.

FLEW, ANTONY: *Equality in Liberty and Justice.* New Brunswick, NJ: Transaction, 1989/2001. xx, 224 p. Paper $24.95.

FORST, RAINER: *Contexts of Justice: Political Philosophy beyond Liberalism and Communitarianism.* JOHN M. M. FARRELL, trans. Berkeley: California UP, 1994/2002. xii, 346 p. Paper $22.50.

FULLER, STEVE: *Thomas Kuhn: A Philosophical History for Our Times.* Chicago: University Press, 2000/2001. xvii, 472 p. Paper $22.50.

GARLAND, ROBERT: *The Greek Way of Death.* Ithaca, NY: Cornell, 1985/1995. xvi, 192 p. Paper $17.95.

GEERTZ, CLIFFORD: *Available Light: Anthropological Reflections on Philosophical Topics.* Princeton: University Press, 2000/2001. xvi, 271 p. Paper $16.95.

HADOT, PIERRE: *The Inner Citadel.* MICHAEL CHASE, trans. Cambridge: Harvard, 1998/2001. xii, 351 p. Paper $18.95.

HAMBURGER, JOSEPH: *John Stuart Mill on Liberty and Control.* Princeton: University Press, 1999/2001. xx, 239 p. Paper $20.95.

HANKINSON, R.J.: *Cause and Explanation in Ancient Greek Thought.* New York: Oxford, 1998/2001. xvi, 499 p. Paper $26.95.

HORN, LAURENCE R.: *A Natural History of Negation.* Stanford: CSLI, 1989/2001. xlvii, 637 p. Paper $30.00.

ISER, WOLFGANG: *The Range of Interpretation.* New York: Columbia, 2000/2001. xv, 206 p. Paper $18.50.

JORDAN, LINDA, and THOMAS G. PAVEL: *The Spell of Language: Poststructuralism and Speculation.* Chicago: University Press, 1989/2001. viii, 178 p. Paper $16.00.

KOŁAKOWSKI, LESZEK: *Husserl and the Search for Certitude.* South Bend, IN: St. Augustine's, 1975/2001. 86 p. Paper $10.00; *Metaphysical Horror.* Chicago: University Press, 1988/2001. vi, 129 p. Paper $16.00; *The Presence of Myth.* ADAM CZERNIAWSKI, trans. Chicago: University Press, 1989/2001. xii, 138 p. Paper $13.00; *Religion, If There Is No God: On God, the Devil, Sin, and Other Worries of the So-Called Philosophy of Religion.* South Bend, IN: St. Augustine's, 1993/2001. 222 p. Paper $18.00.

KUHSE, HELGA, and PETER SINGER, eds.: *A Companion to Bioethics.* Malden, MA: Blackwell, 1999/2001. xv, 512 p. Paper $34.95.

LANGTON, RAE: *Kantian Humility: Our Ignorance of Things in Themselves.* New York: Oxford, 1998/2001. xiv, 232 p. Paper $24.95.

MALCOLM, NORMAN: *Ludwig Wittgenstein, Second Edition.* New York: Oxford, 1984/2001. viii, 136 p. Paper $ 15.95.

MARCEL, GABRIEL: *Creative Fidelity.* ROBERT ROSTHAL, trans. New York: Fordham, 1964/2002. xxxvi, 261 p. Paper $20.00.

MAUDLIN, TIM: *Quantum Non-Locality and Relativity, Second Edition.* Malden, MA: Blackwell, 1994/2002. xi, 281 p. Paper $29.95.

MCDOWELL, JOHN: *Meaning, Knowledge, and Reality.* Cambrige: Harvard, 1998/2002. vii, 462 p. Paper $22.00; *Mind, Value, and Reality.* Cambridge: Harvard, 1998/2002. vii, 400 p. Paper $22.00.

MILL, JOHN STUART: *The Subjection of Women: A Critical Edition.* EDWARD ALEXANDER, ed. New Brunswick, NJ: Transaction, 1869/2001. xxv, 216 p. Paper $29.95.

MILLER, DAVID: *Principles of Social Justice.* Cambridge: Harvard: 1999/2001. xi, 337 p. Paper $18.95.

MORAVCSIK, JULIUS: *Plato and Platonism: Plato's Conception of Appearance and Reality in Ontology, Epistemology, and Ethics, and Its Modern Echoes.* Malden, MA: Blackwell, 1996/2000. ix, 342 p. Paper $31.95.

NAGEL, THOMAS: *The Last Word.* New York: Oxford, 1997/2001. viii, 147 p. Paper $11.95.

NEWTON-SMITH, W. H., ed.: *A Companion to the Philosophy of Science.* Malden, MA: Blackwell, 2000/2001. xvi, 576 p. Paper $34.95.

ORTEGA Y GASSET, JOSE: *Toward a Philosophy of History.* Urbana: Illinois UP, 1941/2002. x, 273 p. Paper $17.95.

PIEPER, JOSEF: *Scholasticism: Personalities and Problems of Medieval Philosophy.* RICHARD WINSTON and CLARA WINSTON, trans. South Bend, IN: St. Augustine's, 1960/2001. 192 p. Paper $13.00.

POPPER, KARL: *The World of Parmenides: Essays on the Presocratic Enlightenment.* ARNE F. PETERSON with JØRGEN MEJER, eds. New York: Routledge, 1998/2001. xvi, 328 p. Paper $19.95.

ROBINSON, HOWARD: *Perception.* New York: Routledge, 1994/2001. xii, 260 p. Paper $27.95.

RUSSELL, BERTRAND: *ABC of Relativity.* New York: Routledge, 1925/2001. xvi, 155 p. Paper $15.95; *The Basic Writings of Bertrand Russell, 1903-1959.* ROBERT E. EGNER and LESTER E. DENONN, eds. New York: Routledge, 1961/2001. 736 p. Paper $26.95; *Common Sense and Nuclear Warfare.* New York: Routledge, 1959/2001. xxvii, 77 p. Paper $14.95.

SALOMÉ, LOU: *Nietzsche.* SIEGFRIED MANDEL, ed. and trans. Urbana: Illinois UP, 1998/2001. lxvi, 168 p. Paper $16.95.

TRAVIS, CHARLES: *The Uses of Sense: Wittgenstein's Philosophy of Language.* New York: Oxford, 1989/2001. xiii, 400 p. Paper $27.95.

## NOTES AND NEWS

### ARTHUR C. DANTO: ART, ACTION, HISTORY

October 3-5, 2002

Miller Theater, Columbia University

The Trustees and Editors of the *Journal of Philosophy*, along with Columbia University's Philosophy Department and School for the Arts, are pleased to announce a conference to honor the work of Arthur C. Danto—as a philosopher, editor, colleague, and art critic. The conference will be held October 3-5, 2002, at Miller Theater, Columbia University, New York City, and will be open to the public.

The speakers will include: Frank Ankersmit, Hans Belting, David Carrier, Stanley Cavell, Donald Davidson, Daniel Dennett, Lydia Goehr, Daniel Herwitz, Gregg Horowitz, Philip Kitcher, Rosalind Krauss, Robert Storr, and Richard Wollheim. There will be four philosophy panels, focusing on the philosophy of art, history, and action, plus two sessions on art history. Finally, there will be a panel discussion on Saturday evening with several artists, curators, and critics addressing Danto's writings and contemporary art.

Please visit our website (www.journalofphilosophy.org or www.jphil.org) for the full program and other related information.

# Young Epistemologist Prize

The **Rutgers Epistemology Conference** announces the third annual prize for the best essay submitted to the Conference by a person with a Ph.D. obtained by the time of submission but not earlier than ten (10) years prior to the date of the conference. Thus, for the Conference on May 2nd/3rd, 2003, the Ph.D. must not have been awarded prior to May, 1993. The prize winning essay will be presented at the annual conference and will be published (along with the invited papers) in Philosophy and Phenomenological Research. The winner of the prize will receive an award of $1,000 plus all travel and lodging expenses connected with the Conference.

*Details*

The essay may be in any area of epistemology. The essay ***must*** be limited to 5,000 words (not including footnotes and bibliography) and submitted on a PC-formatted disk along with three (3) printed copies. The essays will be judged blindly and, thus, the three printed copies (but not the disk copy) must be prepared in such a way that the author's identity is completely masked. *Only* papers submitted in that fashion will be considered. Papers ***must*** be postmarked by December 1, 2002. The winner of the prize will be announced by February 1, 2003. Please include your email address in the cover letter accompanying your paper. By submitting the essay, the author agrees to have it posted on the Rutgers Epistemology Conference Webpage and published in Philosophy and Phenomenological Research.

Please send all papers (along with the disk and a self-addressed postcard that will be returned to entrants acknowledging receipt of the essay) to:

The Young Epistemologist Prize
Philosophy Department
Rutgers University
26 Nichol Ave.
New Brunswick, NJ 08901-2882

For more information about the Rutgers Epistemology Conference and the Young Epistemologist Prize, please refer to:

***http://philosophy.rutgers.edu/epistemology2003/***

## Choosing Character

*Responsibility for Virtue and Vice*

JONATHAN JACOBS

"*Choosing Character* is very well-written. In defending a notion of 'ethical disability' Jonathan Jacobs has much to contribute to current debates on moral responsibility, on the possibility of revision of character, and on the metaethical issues that undergird virtue theory."—Daniel Frank, University of Kentucky. $29.95

## Reasonably Radical

*Deliberative Liberalism and the Politics of Identity*

ANTHONY SIMON LADEN

"Anthony Simon Laden's important book constitutes the philosophical coming-of-age for identity politics. Whereas the claim that liberal theory has tended systematically to silence or ignore the voices of difference has thusfar generated more polemical heat than analytical light, Laden gives the claim its most precise and careful formulation yet."
—Daniel Weinstock, University of Montreal. $39.95

## The Soul of Justice

*Social Bonds and Racial Hubris*

CYNTHIA WILLETT

"Cynthia Willett forcefully argues that a theory of radical democracy needs both a notion of hubris and a notion of social eros. Willett takes us to the heart of tensions between need and independence, autonomy and ultruism, self and others, rights and duties, equality and difference."—Kelly Oliver, State University of New York, Stony Brook. $46.50 cloth, $17.95 paper

# NEW IN PAPER

## Memorabilia

XENOPHON

**Translated and Annotated by Amy L. Bonnette**
**Introduction by Christopher J. Bruell**

An essential text for understanding Socrates, Xenophon's *Memorabilia* is the compelling tribute of an affectionate student to his teacher. Here, Xenophon offers a rare first-hand account of Socrates' life and philosophy. The *Memorabilia* is invaluable both as a work of philosophy in its own right and as a complement to the study of Plato's dialogues. Agora Editions. $17.95

## The Emergent Self

WILLIAM HASKER

"This is the clearest, most systematic and compelling challenge to a materialist philosophy of mind in the current literature. William Hasker develops a series of substantial, well-crafted arguments in the course of articulating a nonreductive account of the self. His book is an outstanding achievement in philosophy of mind."—Charles Taliaferro, St. Olaf College. 1 drawing. $19.95

## Giordano Bruno and Renaissance Science

HILARY GATTI

"[This book] presents a bold argument for Bruno's inclusion not only within the sphere of Renaissance sciences but also within the sphere of science as it is practiced at the present time."
—*The New York Review of Books*. 14 illustrations. $19.95

**The Education of Cyrus**
XENOPHON
Translated and Annotated
by Wayne Ambler

"*The Education of Cyrus* is a major work of classical political philosophy and Wayne Ambler's very impressive English translation is far superior to any other that I have seen."—Peter Ahrensdorf, Davidson College. Agora Editions. 1 table, 1 map, $45.00 cloth, $19.95 paper

**Hegel and the Hermetic Tradition**
GLENN ALEXANDER MAGEE

This book covers Hegel's entire philosophical corpus, showing that his engagement with Hermeticism lasted throughout his entire career and intensified during his final years in Berlin. Viewing Hegel as a Hermetic thinker has implications for a more complete understanding of the modern philosophical tradition, and German idealism in particular. 10 illustrations. $39.95

**A Materialist Metaphysics of the Human Person**
HUD HUDSON

"[This book] is filled with imaginative and memorable examples that enhance its value as a source. Hud Hudson engages with a cluster of difficult issues in ontology, provides an overview of the range of positions on the metaphysics of human persons, thinks through difficult methodological issues, and comes up with original ideas."
—John P. Hawthorne,
Syracuse University. $39.95

**The Political Writings "Selected Aphorisms" and Other Texts**
ALFARABI
Translated and Annotated
by Charles E. Butterworth

Alfarabi was among the first to explore the tensions between the philosophy of classical Greece and that of Islam, as well as of religion generally. This volume presents four of Alfarabi's most important texts, making his political thought available to classicists, medievalists, and scholars of religion and Byzantine and Middle Eastern studies. Agora Editions. $35.00

**"Philosophy of Plato and Aristotle"**
ALFARABI
Translated with an Introduction
by Muhsin Mahdi
Foreword by
Charles E. Butterworth
and Thomas L. Pangle

"[This book] will be of incomparable value to those who are interested in the way in which the major ideas of Plato and Aristotle had been accepted, understood, and developed by the Muslim thinkers within the confines of the Islamic cultural milieu before these great philosophical ideas were introduced in the mainstream of scholastic philosophy in the West."—*Philosophy East and West.* Agora Editions. $14.95 paper

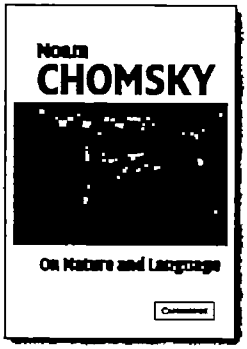
**Cornell University Press**

At bookstores, or call (607) 277-2211 • www.cornellpress.cornell.edu

## Outstanding Scholarship from Cambridge

**A Defense of Abortion**
*David Boonin*
Cambridge Studies in Philosophy and Public Policy
0-521-81701-3 Hardback $65.00
0-521-52035-5 Paperback $20.00

**On Nature and Language**
*Noam Chomsky*
*Adriana Belletti and*
*Luigi Rizzi, Editors*
0-521-81548-7 Hardback $60.00
0-521-01624-X Paperback $20.00



**The Cambridge Companion to Levinas**
*Simon Critchley and*
*Robert Bernasconi, Editors*
Cambridge Companions to Philosophy
0-521-66206-0 Hardback $65.00
0-521-66565-5 Paperback $23.00

**The Transhistorical Image**
Philosophizing Art and its History
*Paul Crowther*
0-521-81114-7 Hardback $55.00

**Meaning, Expression, and Thought**
*Wayne A. Davis*
Cambridge Studies in Philosophy
0-521-55513-2 Hardback $90.00

**Lack of Character**
Personality and Moral Behavior
*John M. Doris*
0-521-63116-5 Hardback $60.00

**Hegel, Nietzsche, and Philosophy**
Thinking Freedom
*Will Dudley*
Modern European Philosophy
0-521-81250-X Hardback $60.00

**Wittgenstein and William James**
*Russell B. Goodman*
0-521-81315-8 Hardback $55.00

**Thought and World**
An Austere Portrayal of Truth, Reference, and Semantic Correspondence
*Christopher Hill*
Cambridge Studies in Philosophy
0-521-81484-7 Hardback $60.00
0-521-89243-0 Paperback $23.00

**Immanuel Kant**
A Biography
*Manfred Kuehn*
0-521-52406-7 Paperback $25.00

**An Essay on the Modern State**
*Christopher W. Morris*
0-521-52407-5 Paperback $23.00

**Autonomy and Trust in Bioethics**
*Onora O'Neill*
0-521-81540-1 Hardback $55.00
0-521-89453-0 Paperback $20.00

**German Philosophy 1760-1860**
The Legacy of Idealism
*Terry Pinkard*
0-521-66326-1 Hardback $65.00
0-521-66381-4 Paperback $23.00

---

*Available in bookstores or from*
CAMBRIDGE
UNIVERSITY PRESS
800-872-7423
us.cambridge.org/philosophy

# Complexities

## Social Studies of Knowledge Practices

JOHN LAW & ANNEMARIE MOL,
EDITORS

Although much recent social science and humanities work has been a revolt against simplification, this volume explores the contrast between simplicity and complexity to reveal that this dichotomy, itself, is too simplistic.

John Law and Annemarie Mol have gathered a distinguished panel of contributors to offer approaches to a theory of complexity, at the same time offering a theoretical introduction to the topic. Indeed, they examine not only ways of relating to complexity but complexity in practice.

With a focus on such concepts as multiplicity, partial connections, and ebbs and flows, the collection includes narratives from Kenya, Great Britain, Papua New Guinea, the Netherlands, France, and the meetings of the European Commission, written by anthropologists, economists, philosophers, psychologists, sociologists, and scholars of science, technology, and society.

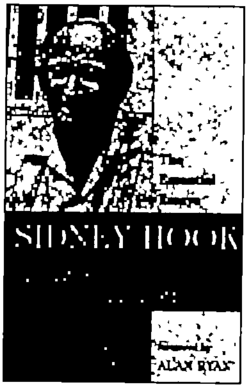
304 pages, 4 b&w photos, 2 tables, 9 figures, paper $19.95
*Science and Cultural Theory*

DUKE UNIVERSITY PRESS
www.dukeupress.edu toll-free 1-888-651-0122

*New Philosophy Titles from . . .*

Prometheus Books

**SIDNEY HOOK ON PRAGMATISM, DEMOCRACY, AND FREEDOM**
The Essential Essays
Edited by Robert B. Talisse
and Robert Tempio
*Foreword by Alan Ryan*
420 pp • ISBN 1-59102-022-0 • HC $32



**THE SCIENCE WARS**
Debating Scientific Knowledge
and Technology
Edited by Keith Parson
300 pp • ISBN 1-57392-994-8 • PB $21

**PRAGMATIC NATURALISM AND REALISM**
Edited by John R. Shook
325 pp • ISBN 1-57392-997-2 • HC $32

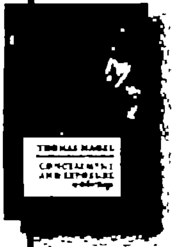
*. . . and*

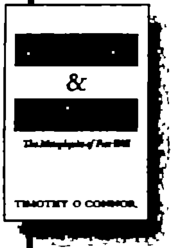
Humanity Books

**NATIVE AMERICAN WORLDVIEWS**
An Introduction
Jerry H. Gill
*"I found his book to be a treasure. Concepts that I had been trying to to convey to students in bits and pieces over the course of a semester are brought together in a coherent overview which differs from simple historical, archeological, or geographical presentations of Native Americans."*
—Theresa A. Vaughan, Ph.D.,
Assistant Professor of Humanities and Philosophy,
University of Central Oklahoma
260 pp • ISBN 1-59102-051-4 • PB $25

**GLOBALIZATION AND JUSTICE**
Kai Nielsen
300 pp • ISBN 1-59102-054-9 • HC $39

For exam copies contact editorial@prometheusbooks.com
Call toll free 800-421-0351 • Fax 716-691-0137 • www.prometheusbooks.com





New from Oxford

# REQUIRED READING

*New in Paperback!*

## THE ARCHITECTURE OF REASON

**Robert Audi**

"A wonderful book. It systematically tackles a fundamental, pervasive philosophical issue in lucid prose, and it makes a powerful case for an extremely attractive unity."—Alfred Mele, *Times Literary Supplement*

"One of the most richly elaborated and widely-ranging rationality theories to date."
—Stephen W. Ball, *Utilitas*

304 pp. $19.95

## CONCEALMENT AND EXPOSURE

**And Other Essays**

**Thomas Nagel**

Widely recognized as one of the top American philosophers working today, this volume is a collection of his most recent critical essays and reviews.

288 pp. $24.95 cloth

## RADICAL INTERPRETATION AND INDETERMINACY

**Timothy McCarthy**

The major theme in McCarthy's study is that a relatively modest set of interpretive principles, properly applied, can serve to resolve the major indeterminacies of interpretation.

272 pp. $45.00 cloth

## HUME'S REASON

**David Owen**

"Owen's scholarship is meticulous and his attempt to understand Hume's historical context commendable."
—*Review of Metaphysics*

248 pp. $19.95 paperback

## INTRODUCTION TO A PHILOSOPHY OF MUSIC

**Peter Kivy**

Philosophy of music has flourished in the last thirty years, with great advances made in the understanding of the nature of music and its aesthetic. Anyone interested in music will find this a stimulating introduction to some fascinating questions and ideas.

244 pp. $18.95 paperback

## PERSONS AND CAUSES

**The Metaphysics of Free Will**

**Timothy O'Connor**

"The book is intelligent throughout. O'Connor is unafraid to defend an unfashionable view, and to do so in a bold and imaginative way."
—*MIND*

136 pp $19.95 paperback

*New in Paperback!*

## MEANING, UNDERSTANDING, AND PRACTICE

**Philosophical Essays**

**Barry Stroud**

This collection offers penetrating studies of meaning, understanding, necessity, and the intentionality of thought. Stroud's helpful introduction draws out the recurring themes he pursues in the essays and explains how his ideas and aims have developed over the years.

252 pp. $19.95

Prices are subject to change and [illegible]

New from **Oxford**

# REQUIRED READING

### THE QUEST FOR REALITY

**Subjectivism and the Metaphysics of Colour**

**Barry Stroud**

"The book's leisurely, discursive, and explicitly undogmatic treatment makes for a fascinating and pleasant read."

—*Review of Metaphysics*

256 pp $19.95 paperback

### NIETZSCHE'S SYSTEM

**John Richardson**

"Noteworthy for its engagement with rival interpretations of Nietzsche, and for its scrupulous attention to the many recalcitrant passages in his writings. A useful and provocative contribution to the understanding and evaluation of Nietzsche's thought."—*Ethics*

328 pp. $24.95 paperback

### PLATO'S RECEPTION OF PARMENIDES

**John A. Palmer**

"This excellent and extensive survey is valuable reading for scholars and graduate students interested in Plato's metaphysics and epistemology."

—*Choice*

312 pp. $19.95 paperback

### ESSAYS IN THE METAPHYSICS OF MODALITY

**Alvin Plantinga, edited by Matthew Davidson**

These essays, dating from the late 1960's to the present, chronicle the development of Plantinga's thoughts about some of the most fundamental issues in metaphysics.

304 pp. $24.95 paperback, $60.00 cloth

### ENGAGING REASON

**On the Theory of Value and Action**

**Joseph Raz**

A penetrating exploration of the interdependence of value, reason, and the will. The essays illuminate a wide range of questions concerning fundamental aspects of human thought and action.

344 pp. $19.95 paperback

### PHILOSOPHICAL RELATIVITY

**Peter Unger**

In this short volume (first published in 1984), Peter Unger questions the objective answers that have been given to traditional problems in philosophy. Written with unusual clarity, this book is provocative, highly readable, and ambitious in scope.

144 pp. $19.95

***New in Paperback!***

### CAUSING ACTIONS

**Paul M. Pietroski**

Presenting an original philosophical theory of actions and their mental causes, Pietroski argues that thoughts and deeds are in fact distinct from, though dependent on, underlying biochemical processes.

288 pp. $19.95

OXFORD
UNIVERSITY PRESS

Prices are subject to change and apply in the US. To order, call 1-800-451-7556. In Canada, 1-800-387-8020. Visit our web site at **www.oup-usa.org**

# THE JOURNAL OF PHILOSOPHY

25\11\02

VOLUME XCIX, NUMBER 10
OCTOBER 2002

*page*
499 *Robust Nonreductive Materialism* Derk Pereboom
532 *Trust out of Distrust* Edna Ullmann-Margalit

549 NEW BOOKS
552 NOTES AND NEWS

***Published by*** **The Journal of Philosophy, Inc.**

# THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

*Editors*: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: Michael Kelly.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to Michael Kelly, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX. (212) 932-3721.
You may also visit our website at: www.journalofphilosophy.org

# THE JOURNAL OF PHILOSOPHY

2003

## SUBSCRIPTIONS (12 issues)

| | |
|---|---|
| Individuals | $35.00 |
| Libraries and Institutions | $75.00 |
| Students, retired/unemployed philosophers | $20.00 |
| Postage outside the U.S. | $15.00 |

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Ephrata, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL'S terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by writing to us.

POSTMASTER. Periodical postage paid at New York, NY, and other mailing offices.
POSTMASTER: Send address changes to the *Journal of Philosophy* at MC 4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.
© Copyright 2002 by the Journal of Philosophy, Inc. ISSN 0022–362X

THE JOURNAL OF PHILOSOPHY
VOLUME XCIX, NO. 10, OCTOBER 2002

25/11/02

# ROBUST NONREDUCTIVE MATERIALISM*

Nonreductive materialism about the mental has been put on the defensive by a series of well-developed arguments against its central claims. Four of these challenges, each of which has been advanced by Jaegwon Kim, are especially prominent: the argument from explanatory exclusion against irreducibly mental causal powers; the contention that the nonreductive view is indistinguishable from the emergentism of Samuel Alexander and C. Lloyd Morgan, a position widely held to be metaphysically extravagant; the claim that the functionalism typically endorsed by nonreductive materialists is incompatible with irreducibly mental causal powers; and the argument that, if mental state types are multiply realizable, they cannot be genuinely scientific kinds, for then they will be only as weakly projectible as the wild disjunction of their possible realizations. This last challenge is representative of a growing skepticism about arguments against reductionism from multiple realizability.

I shall first examine whether nonreductive materialism can finesse the explanatory exclusion problem. Subsequently, I shall argue that there are significant differences between the controversial sort of emergentism and nonreductive materialism, and that a nonreductive materialist need not be emergentist in this sense. I shall then contend that a position according to which mental states instantiate irreducibly mental causal powers—the key feature of what I shall call *robust nonreductive materialism*—indeed cannot be functionalist, but that there is a nonfunctionalist account of mental states to which the nonreductivist can turn. I shall close by examining doubts that have been raised about arguments from multiple realizability against re-

---

*Thanks to Ned Block, David Christensen, Hilary Kornblith, and Mark Moyer for valuable comments and discussion.

0022-362X/02/9910/499–531

ductionism, concluding that the nonreductive view can withstand these doubts.

## I AVOIDING EXPLANATORY EXCLUSION

Robust nonreductive materialism, as I conceive it, is a view about specifically psychological explanations, states, and causal powers, although it easily generalizes to other levels of explanation. In this view, an event such as Mary's buying ice cream (M2) will have a psychological explanation in terms of a complex of mental states—beliefs and desires she has (M1). Each of M1 and M2 will be wholly constituted of microphysical events (P1 and P2 respectively), and there will be a microphysical explanation of P2 in terms of P1. The explanation of M2 by M1 will not reduce to the explanation of P2 by P1 (and, likewise, mutatis mutandis for events at various other levels of description). Underlying the irreducibility of this explanation is the fact that M1 is not type identical with P1, and that M2 is not type identical with P2. More fundamentally yet, the psychological explanation appeals to the irreducibly mental causal powers of M1 to account for M2, while the microphysical explanation appeals to microphysical causal powers of P1 to account for P2. Accordingly, the causal powers of M1 will not be type identical with those of P1, and those of M2 will not be type identical with those of P2. But neither will a corresponding token-identity thesis for these causal powers hold. For if it did, then the causal powers to which the psychological explanation refers would in the last analysis, in fact, be microphysical. Psychological explanations might then presume a classification that clusters microphysical causal powers in a way distinct from how microphysics sorts them, but this would not compromise the microphysical status of those causal powers. Hence, robust nonreductive materialism affirms various token-diversity claims for mental causal powers. The token mental causal powers of M1 and M2 will not be identical with the token microphysical causal powers of P1 and P2, nor with the token neural causal powers of the neural states N1 and N2 that constitute M1 and M2, nor with token causal powers at any other level of description more basic than the neural.[1]

Furthermore, in my version of this robust nonreductive conception, there will be a microphysical explanation for P2 which appeals to the microphysical causal powers of P1, and at the same time P2 (together with any requisite relational features) will be sufficient for M2. Consequently, there will be a microphysical causal explanation for M2 that appeals to the microphysical causal powers of P1 (given a constitutional account of M2 in terms of P2). For since one standard way of explaining an event causally is to cite a causal power whose activation is a sufficient cause of this event, citing the causal powers of P1 yields a causal explanation of M2. In another prominent nonreductive model, deriving from Hilary Putnam,[2] there exists no genuine microphysical explanation for the action, only a psychological one. I am strongly persuaded by the line of reasoning just presented against this view. Accordingly, I shall set the alternative model aside in this discussion.

---

[1] In *Mind in a Physical World* (Cambridge: MIT, 1998), pp. 67-87, Kim discusses several nonreductive views that are not robust in this sense.

Familiarly, the position under scrutiny gives rise to a pressing question: What is the relationship between the microphysical and psychological explanations for M2? In particular, given that both sorts of explanation refer to causal powers, what is the relationship between the causal powers to which the microphysical explanation appeals and those to which the psychological explanation appeals? Here is where Kim's challenge from causal or explanatory exclusion enters in.[3] As we have seen, if a microphysical account yields a causal explanation of the microphysical constitution of M2, then it will provide a causal explanation of M2 itself. What room is then left for a distinct psychological causal explanation of this action? Kim argues that it is implausible that the psychological explanation appeals to causal powers whose activation is sufficient for the event to occur, and at the same time the microphysical explanation appeals to distinct causal powers also sufficient for the event to occur, and that as a result the event is overdetermined. But it is also implausible that each of these distinct sets of causal powers yields a partial cause of the event, and that each by itself would be insufficient for the event to occur.

By the solution to this problem which Kim develops, real causal powers exist at the microphysical level, and so the microphysical explanations refer to real microphysical causal powers. Only if psychological explanations in some sense reduce to microphysical explanations does it turn out that the psychological explanations also appeal to real causal powers—these causal powers will then ultimately be microphysical. Psychological explanations that do not reduce to microphysical explanations will fail to refer to causal powers, and thus will have some lesser status—such explanations might express regularities without at the same time referring to causal powers. This move solves the exclusion problem because if the causal powers to which the psychological explanation appeals are identical with those to which the microphysical explanation appeals, then there will be no genuine competition between explanations, and if the psychological explanations do not refer to causal powers at all, there will be no competition either. But this solution, which Kim believes is the only possible solution to the problem he raises, would rule out any nonreductive view about mental causal powers.

---

[2] "Language and Reality," in *Philosophical Papers, Volume 2* (New York: Cambridge, 1975), pp. 272-90, at p. 278; (this paper was delivered as a Machette Lecture at Princeton, May 22, 1974). Elliott Sober makes a strong case for doubting the applicability of this alternative—"The Multiple Realizability Argument against Reductionism," *Philosophy of Science*, LXVI (1999): 542-64.

[3] See, for example, Kim, "The Myth of Nonreductive Materialism," in his *Supervenience and Mind* (New York: Cambridge, 1993), pp. 265-84, at pp. 280-82; this article was first published in *Proceedings and Addresses of the American Philosophical Association*, LXIII (1989): 31-47. Stephen Yablo provides a historical bibliography of the causal/explanatory exclusion argument—"Mental Causation," *The Philosophical Review*, CI (1992): 245-80, at p. 247, note 5.

In Kim's conception, any token causal powers of a higher-level property at a time will be identical with some token (micro)physical causal powers. He applies this view to mental properties as:

> [The causal inheritance principle:] If mental property *M* is realized in a system at *t* in virtue of physical realization base *P*, the causal powers of *this instance of M* are identical with the causal powers of *P*.[4]

Kim contends that rejecting this principle would be tantamount to accepting "causal powers that magically emerge at a higher level and of which there is no accounting in terms of lower-level causal powers and nomic connections" (*ibid.*, pp. 326-27). By the causal inheritance principle, there would be no token causal powers distinct from token microphysical causal powers, which would preclude any robust nonreductive materialism. Higher-level kinds and explanations would at best group token microphysical causal powers in a way that does not correspond to the classifications of microphysics itself.[5] Such a classification might be of value for prediction. But there would remain no sense in which there exist causal powers that are not microphysical.

Let us first examine the causal inheritance principle, and then return to the explanatory exclusion issue. Is the causal inheritance principle true? And if it is false, is Kim right to suppose that magical emergentism follows? I think that the answer to both of these questions is "no." First, a respectable case can be made that higher-level token entities are typically not identical with their realization bases. The ship of Theseus is not identical with its current token microphysical realization base, for it would have been the same token ship had the token microphysical realization been slightly different, and it will be the same ship when this microphysical realization in fact changes—the ship is in this sense *token multiply realizable*. True, there is some notion of sameness by which the ship is the same thing as its current microphysical realization—a notion that abstracts from any temporally extrinsic or modal properties.[6] But it is sufficient for absolute nonidentity that $A$ and $B$ differ in their temporally extrinsic or modal properties. The same sort of argument can be run for token mental entities.[7] Is token mental state $M$ identical with $P$, its actual token microphysical realization base? Suppose that $M$ is realized by a complex neural state. It is possible for $M$ to be realized differently only in that a few neural pathways are used that are token distinct from those actually engaged. We need not rule at this point on whether the actual neural realization is token identical with this alternative—it might well be. (Just as the ship of Theseus would retain its identity supposing the replacement of a few of its planks, so it would seem that a token neural state would retain its identity given the replacement of a few of its neural pathways—more on this later.) But it is evident that this alternative neural realization is itself realized by a microphysical state $P^*$ that is token distinct from $P$. It is therefore possible for $M$ to be realized by a microphysical state not identical with $P$, and thus $M$ is not identical with $P$. But, furthermore, this reflection would also undermine a token-identity claim for mental causal powers—should they exist—and their underlying microphysical causal powers. For if the token microphysical realization of $M$ had been different, its token microphysical causal powers would also have been different. We therefore have good reason to suppose that any token mental causal powers of $M$ would not be identical with the token microphysical causal powers of its realization.

---

[4] Kim, "Multiple Realization and the Metaphysics of Reduction," in his *Supervenience and Mind*, pp. 309-335, at p. 326 (italics added ); this article was first published in *Philosophy and Phenomenological Research*, LII (1992): 1-26.

[5] Terence Horgan, "Kim on Mental Causation and Causal Exclusion," *Philosophical Perspectives*, XI (1997): 165-84, at p. 179. See also Kim's discussion of this move in *Mind in a Physical World*, pp. 67-72.

Still, there would be a sense in which the token causal powers of $M$ would be "nothing over and above" the token causal powers of

---

[6] See Mark Moyer, *A Semantic Approach to Material Constitution*, unpublished Ph.D. Dissertation, Department of Philosophy, Rutgers University, New Brunswick, New Jersey, 2002.

[7] Cf. Saul Kripke, *Naming and Necessity* (Cambridge: Harvard, 1980), pp 144-48; Pereboom and Hilary Kornblith, "The Metaphysics of Irreducibility," *Philosophical Studies*, LXIII (1991)· 125-45, at pp. 131-32; and Lynne Rudder Baker, *Explaining Attitudes. A Practical Approach to the Mind* (New York: Cambridge, 1995), pp. 9-10.

*P*—*M*'s causal powers would nevertheless be "absorbed" or "swallowed up" by *P*'s causal powers.[8] But there are two importantly distinct modes of this sort of absorption: identity and constitution without identity. And if there were essentially mental causal powers (that are physically realized), the relation of any one such token to its microphysical realization base would be the second and not the first.

On this picture, a token mental state would have the mental causal powers it does ultimately by virtue of the token microphysical states of which it is constituted (setting aside any fundamentally relational causal powers). For this reason, it makes sense to say that token mental causal powers are wholly constituted by token microphysical causal powers. More generally, as Hilary Kornblith and I have contended,

> [Token causal power constitution:] The causal powers of a token of kind *F* are constituted of the causal powers of a token of kind *G* just in case the token of kind *F* has the causal powers it does in virtue of its being constituted of a token of kind *G* (*op. cit.*, p. 131, italics added).

This nonreductive view endorses a weaker but nevertheless plausible version of the causal inheritance principle:

> [The weaker causal inheritance principle:] If mental property *M* is realized in a system at *t* in virtue of physical realization base *P*, the causal powers of this instance of *M* are wholly constituted by the causal powers of *P*.

Moreover, on this view there will be a significant degree to which causal powers of higher-level tokens could be explained in terms of the causal powers of their microphysical constituents—limited by the extent to which the causal powers of the higher-level tokens are relational.[9] Furthermore, correlated with the possibility of this sort of constitutional explanation is the fact that the existence and nature of token higher-level causal powers would be predictable in principle from their microphysical constituents together with the laws governing them. This predictability would again be limited by the extent to which the higher-level powers are relational, but, as I shall argue in the next section, nothing else functions to impede it.

---

[8] Cf. John Heil, "Multiple Realizability," *American Philosophical Quarterly*, XXXVI (1999): 189-208.

[9] Because it allows for this sort of constitutional explanation, nonreductive materialism does not have the consequence that information about the brain is of little or no relevance to understanding psychological processes (cf. Pereboom and Kornblith, pp. 140-42). William Bechtel and Jennifer Mundale express the concern that nonreductive materialism will have this sort of result—"Multiple Realizability Revisited: Linking Cognitive and Neural States," *Philosophy of Science*, LXVI (1999): 175-207, at p. 176.

Just as Kim claims that no competition between explanations arises in the case of reduction and identity, I propose that no competition arises in the case of mere constitution either.[10] For if the token of a higher-level causal power is currently wholly constituted by a complex of microphysical causal powers, there are two sets of causal powers at work which are constituted from precisely the same stuff (supposing that the most basic microphysical entities are constituted of themselves), and in this sense we might say that they *coincide constitutionally*.[11] The fact that they now coincide in this way might tempt one to suppose that these causal powers are token identical, but, as we have just seen, there is a good argument that they are not. And because it is possible for there to be wholly constitutionally coinciding causal powers that are not even token identical, it is possible that there be two causal explanations for one event that do not exclude each other and at the same time do not reduce to a single explanation.[12]

---

[10] Here, I develop an argument I suggested in *Living without Free Will* (New York: Cambridge, 2001), pp. 77-78.

[11] Conceivably, two space ships might be made of such extraordinary material that they can fly through each other, for a moment wholly coinciding spatially. Intuitively, their causal powers might be explanatorily competitive. But it is not conceivable (at least to me) that two entities constituted of exactly the same stuff—two entities that are constitutionally coincident—might in this way be explanatorily competitive. Thanks to Mark Moyer for this point.

[12] Barry Loewer argues that Kim's reasons for concern about overdetermination apply only to multiple determination by independent causes (like two assassins acting independently) but not to cases of overdetermination in which the causes are metaphysically connected (in his review of Kim's *Mind in a Physical World*, this JOURNAL, XCVIII, 6 (June 2001): 315-24). I endorse the substance of Loewer's criticism. As a terminological matter, however, it is not obvious to me that cases in which the causes are metaphysically connected in the way that M1 and P1 are should be classified as cases of overdetermination. The paradigm cases of overdetermination are those in which the causes are independent of one another (in a sense that one might want to spell out further) and do not include those in which the causal processes coincide constitutionally. Kim surely had in mind such paradigm cases, and he would be right to suggest that the mental/physical case does not fit this model. On the one hand, one might argue that due to such constitutional coincidence, it is best to say that there is no multiple determination, and thus no overdetermination, of M2 by P1 and M1, but rather a single determination, just as there would be if the mental and microphysical causal powers (and their activations) were identical. Or one might contend that because the causes are not identical, the case is best classified as one of overdetermination. I am neutral on this question—it seems clear that the important philosophical considerations are independent of any resolution of this terminological issue. Thomas Crisp and Ted Warfield also propose that Kim too hastily dismisses the overdetermination solution—"Kim's Master Argument," *Noûs*, XXXV (2001): 304-16.

If identity and not just constitutional coincidence were necessary for explanatory noncompetition, then there would be features required for noncompetition that identity has and current constitutional coincidence lacks. The candidate features would arguably be constitutional coincidence at all other times, and constitutional coincidence at all other possible worlds, even now. But it is difficult to see how the token causal powers' constitutional noncoincidence at some past time, or at some future time, or their merely possible constitutional noncoincidence even now would introduce explanatory competition, while actual current constitutional coincidence in absence of any features of this sort (that is, identity) would guarantee noncompetition. Suppose that my current token mental state $M$ actually constitutionally coincides with token microphysical state $P$. Let us agree with Kim, that if $M$ were identical with $P$, and if their causal powers were identical, there would be no explanatory competition. If mere constitutional coincidence without identity resulted in explanatory competition, that would have to be because at some time in the past or in the future, or at some other possible world even now, $M$ and $P$ and their causal powers are constitutionally noncoincident. Suppose that $M$ would still exist even if a few neural pathways in its neural realization were token distinct from what they actually are. These neural changes would render $M$'s microphysical realization base distinct from $P$, and thus $M$ and $P$ would be constitutionally noncoincident in some other possible world, and, similarly, mutatis mutandis for their causal powers. How could a possibility of this sort introduce explanatory competition? It would appear that actual current constitutional coincidence alone is relevant to securing noncompetition, and thus for this purpose constitutional coincidence without identity would serve as well as identity.[13] Consequently, it would appear that available to the nonreductivist is a solution to the exclusion problem no less adequate than Kim's own.

---

[13] The proposal that the causal powers of $M$ and $P$ do not compete because they coincide constitutionally is one of a broader class of proposals which also includes Yablo's suggestion that the causal powers of $M$ and $P$ do not compete because they are related as determinable and determinate—"Mental Causation," and "Wide Causation," *Philosophical Perspectives*, XI (1997): 251-81. His example of properties that stand in this relation are *red* and *scarlet*—*red* is a determinable of which *scarlet* is a determinate. It is clear that *red* and *scarlet* will not compete explanatorily. The relation between these two properties is more intimate than constitutional coincidence, and so if the relation between the causal powers of $M$ and $P$ could be shown to fit this model, there may be a respect in which we might then have a more satisfying solution yet to Kim's problem. But still, given that the causal powers of $M$ and those of $P$ coincide constitutionally, even if $M$'s are not a determinable of which $P$'s are the determinate, it is hard to see why the solution we now have to the explanatory exclusion problem is not just as adequate as Kim's.

## II DISTINGUISHING THE NONREDUCTIVE VIEW FROM EMERGENTISM

Kim contends that nonreductive materialism is committed to emergentism: "The fading away of reductionism and the enthronement of nonreductive materialism as the new orthodoxy simply amount to the resurgence of emergentism—not all of its sometimes quaint and quirky ideas but its core ontological and methodological doctrines."[14] In his analysis, emergentism claims a distinction between two sorts of higher-level properties, *resultant* and *emergent*, that arise from the basal conditions of physical systems (*ibid.*, pp. 6-7). The basal conditions of a physical system comprise (i) the basic particles that constitute the physical system, (ii) all the intrinsic properties of these particles, and (iii) the relations that configure these particles into a structure. The higher-level properties that are merely resultant are simply and straightforwardly calculated and theoretically predictable from the facts about its basal conditions—which presumably include the laws that govern the basal conditions—while those that are emergent cannot be calculated and predicted. (Note that the variety of "predictability" at issue is not the possibility of discerning the future given current conditions, but rather the possibility of determining from an entity's realization base what its concurrent higher-level properties are—it is in this sense synchronic and not diachronic predictability.) Theoretical predictability contrasts with inductive predictability. Having regularly witnessed that an emergent property is realized by particular basal conditions, we would be able to predict this relationship, but this sort of inductive predictability is not at issue. Rather, emergentists maintain that knowledge of the basal conditions alone, no matter how complete, will not suffice to yield a prediction of an emergent property (*ibid.*, p. 8). Emergentism of this sort (sometimes called *strong* emergentism) is often regarded as having next to no scientific credibility.[15] I have claimed, however, that on the nonreductive materialist view higher-level properties are, in

[14] "Making Sense of Emergence," *Philosophical Studies*, xcv (1999): 3-36, at p. 5.

[15] Brian McLaughlin, "The Rise and Fall of British Emergentism," in A. Beckermann, H. Flohr, and Kim, eds., *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism* (New York: de Gruyter, 1992); Mark Bedau, "Weak Emergence," *Philosophical Perspectives*, xi (1997): 375-99, at pp. 376-77; Alexander Rueger, "Physical Emergence, Diachronic and Synchronic," *Synthese*, cxxxiv (2000): 297-322, at pp. 317-18. For dissenting voices, see Timothy O'Connor, "Emergent Properties," *American Philosophical Quarterly*, xxxi (1994): 91-104; and John Dupré, "The Solution to the Problem of the Freedom of the Will," *Philosophical Perspectives*, x (1996): 385-402. Strong emergence is sometimes contrasted with weak emergence (for example, Bedau, Rueger), a notion we can set aside for present purposes. Randolph Clarke nicely characterizes emergentism and argues that the nonreductivist can avoid it—"Nonreductive Physicalism and the Causal Powers of the Mental," *Erkenntnis*, li (1999): 295-322.

fact, predictable from basal conditions (except insofar as higher-order properties are relational[16]). If this is correct, the nonreductive view is not committed to (this often-dismissed variety of) emergentism. We shall revisit the predictability issue shortly.

In Kim's analysis, a further characteristic of emergent properties is that they cannot be reductively explained in terms of the physical basal conditions. Kim rejects Ernest Nagel's[17] "bridge law" conception of reductive explanation in favor of a model that first functionalizes the higher-level property to be reduced, and then identifies that property with the realizer of that functionalization in the physical base. Here, one might be interested in finding a particular realizer for a particular instance of the higher-level property, or one might be interested in finding the general realizer for some species or structure type. Finally, one finds a theory at the level of the physical base that explains how the realizers of the higher-level property can instantiate its functional specification.[18] In this conception, a property turns out to be emergent if it cannot be functionalized, or else if no realizer in the physical base can be found for specific instances of the property or for species or structure types—no entity in the physical base for which there is a theory that can explain how it can realize that property's functional specification. One should note here, that if there is to be a debate about whether nonreductivists might avoid emergentism, this irreducibility condition should at least initially be viewed as necessary but not sufficient for an emergent property.

Emergentism also endorses downward causation—it claims that higher-level states can have lower-level effects. (Kim raises a serious difficulty for the synchronic reflexive version of downward causation; here we shall assume the diachronic variety (*ibid.*, pp. 28-31).) As applied to the topic at hand, emergentism asserts that a mental event can cause a microphysical event. According to Kim, the problem with this claim again derives from causal exclusionary considerations. Suppose mental event M1 causes microphysical event P2. M1 will be constituted of some microphysical event, P1, and M1 and P1 will compete as the cause of P2, and P1 will ultimately win out. Only by identifying M1 and P1 can M1's status as cause be salvaged.

---

[16] Whether a large piece of rock is a planet (*being a planet* is a relational property) is not predictable from its basal conditions, at least if the particles that constitute the system are only those which constitute the piece of rock itself. But, intuitively, this does not make *being a planet* an emergent property.

[17] *The Structure of Science* (New York: Harcourt, Brace, and World, 1961), chapter 11.

[18] "Making Sense of Emergence," pp. 10-11.

To my mind, nonreductive materialism indeed countenances downward causation of this sort, and it can legitimately do so because one can reasonably hold that, if M1 is wholly constituted of P1, M1 and P1 will not compete as causes of P2. But its allowing for downward causation is not by itself sufficient to render the nonreductive position in any sense magical or radical. It is, for instance, wholly compatible with the theoretical predictability of higher-level properties from basal conditions. In my view, an emergentist's endorsement of downward causation would be radical if it specified that emergent properties could effect changes in the laws that govern the microphysical level independently of any emergent properties (henceforth *ordinary* microphysical laws). Supposing that M1 were such an emergent property, M1 could then cause P2 in such a way that P2 is no longer governed by the ordinary microphysical laws, but instead by laws that take into account the special characteristics of the emergent properties (or no laws at all). Elsewhere,[19] I have argued that given a materialist metaphysics, agent-causal libertarianism would be committed to the position that the agent-causal power is law-altering in this sense. But nothing essential to nonreductive materialism entails this radical variety of downward causation.

As Clarke explains (*op. cit.*, p. 309), this proposed feature of emergent properties would preclude their theoretical predictability. For an emergent property's capacity for altering the ordinary microphysical laws would not be predictable from a microphysical base given knowledge of only these laws. And supposing that the capacity for altering these laws is at least part of what provides the emergent property with its special nature, the property itself would then not be predictable from the microphysical base given knowledge of only these ordinary laws. But as Clarke points out, no feature of the nonreductive model per se renders higher-level properties any less theoretically predictable than they would be on a reductive model. True, according to the nonreductive view, knowledge of the basal conditions of some entity may not facilitate full predictability of its higher-level relational features, but this is also the case for reductionism. In each model, holding higher-level relational conditions fixed, a particular set of basal conditions will necessitate the same unique higher-level properties. An emergentist might also accept this, but what is crucial is that the nonreductivist is no more beholden than is the reductionist to some factor that threatens to inhibit theoretical predictability, such as the power of higher-level properties to alter the ordinary microphysical laws. There is a difference in that the nonreductive view claims that there are higher-level causal powers that are multiply realizable, and knowledge of the higher-level features of causal powers alone would in such cases not allow the prediction of actual basal conditions. But this fact does not imperil the predictability of higher-level causal powers from basal conditions. Consequently, knowledge of the basal conditions can provide for exactly the same capacity for predicting higher-level properties in each model. And thus, what according to Kim is the defining feature of emergentism turns out not to be an essential characteristic of nonreductive materialism.

---

[19] *Living without Free Will*, pp. 69-74.

Kim further argues that, if the physical base is nomologically sufficient for mental properties, then explanations in terms of mental states will be dispensible. In his example, emergent mental state *M* causes *M**, *M*'s base is physical state *P* and *M**'s base is physical state *P**. The point he is making works as well if *M* is irreducible rather than emergent.

> Now we are faced with *P*'s threat to preempt *M*'s status as a cause of *P** (and hence of *M**). For if causation is understood as nomological (law-based) sufficiency, *P*, as *M*'s emergence base, is nomologically sufficient for it, and *M*, as *P**'s cause, is nomologically sufficient for *P**. Hence *P* is nomologically sufficient for *P** and hence qualifies as its cause. The same conclusion follows if causation is understood in terms of counterfactuals—roughly, as a condition without which the effect would not have occurred.... This appears to make the emergent property *M* otiose and dispensable as a cause of *P**; it seems that we can explain the occurrence of *P** simply in terms of *P*, without invoking *M* at all.[20]

On my nonreductive proposal it is indeed true that *P* is nomologically sufficient for *P**, and that *P** might well be asymmetrically counterfactually dependent on *P*, and that given these analyses of causation, *P* causes *P**. In fact, it seems reasonable to demand that any analysis of causation should count *P* as causing *P**. It also seems right to agree with Kim that we can therefore explain the occurrence of *P** simply in terms of *P*, without invoking *M* as a cause at all. Indeed, we can also explain the occurrence of *M** simply in terms of *P*, without invoking *M* as a cause at all. But does any of this make *M* dispensable as a cause of *P**, and can we do away with an explanation of *P** in terms of *M* as a cause?

---

[20] "Making Sense of Emergence," p. 32, italics added.

I do not believe so. First of all, I have argued that token mental causal powers would not be identical with token microphysical causal powers. Suppose that M instantiates an irreducibly mental causal power, and that the activation of this irreducibly mental causal power brings about *P**. Accordingly, *M* will cause *P**, and *M*'s causing *P** will not be identical with *P*'s causing *P**, and thus there is a significant sense in which *M* is indispensable as a cause of *P**. Given a conception according to which explanations track and express the causal relations in which causal powers participate, the explanation of *P** in terms of *M* will be *bona fide* and indispensible. If one were to claim instead that genuine explanations account for phenomena only in terms of the most fundamental conditions that cause them, then explanations in terms of mental causal powers might well be dispensable, while only microphysical explanations would survive. But this is not a notion of explanation that we are constrained to accept.

Contrary to Kim's claims, the nonreductive materialist is not forced to affirm higher-level causal powers of the (strongly) emergent type. In short, nonreductive materialism need not agree that mental states have the power to produce deviations from the ordinary microphysical laws, and hence it can avoid this potential hindrance to theoretical predictability of mental states from basal conditions. And no other impediment to this sort of predictability that is not also a feature of reductionism appears on the horizon.

## III. A NONFUNCTIONALIST ACCOUNT

By way of protest against Kim's species- or structure-specific reductionism, Ned Block[21] once asked: "What is common to the pains of dogs and people (and all other species) in virtue of which they are pains" (*ibid.*, pp. 178-79)? And as Kornblith and I argued,

> ...even if there is a single type of physical state that normally realizes pain in each type of organism, or in each structure type, this does not show that pain, *as a type of mental state*, is reducible to physical states. Reduction, in the present debate, must be understood as reduction of types, since the primary object of reductive strategies are explanations and theories, and explanations and theories quantify over types... (*op. cit.*, p. 135)

In reply to this objection, Kim points out that nonreductive materialists typically argue from a functionalist perspective, and that functionalists characterize mental states solely in terms of purely relational or extrinsic features of those states. Indeed, functionalism identifies mental state types with type-level dispositions to cause mental states and behavioral outputs given perceptual inputs and mental states—with the understanding that these dispositions are purely relational, that they are to be analyzed in terms of causal relations to perceptual inputs, behavioral outputs, and other mental states, and no intrinsic mental components. Functionalists claim that what all pains would have in common, by virtue of which they are all pains, is a pattern of such relations described by some functional specification—call it *H*. Kim then argues that in providing an answer to Block's question, the local reductionist—the one who opts for species- or structure-specific reductionism—is no worse off than the functionalist. Both are committed to the claim that there is no nonrelational or intrinsic property of pain that all pains have in common, and both can specify only shared relational properties:

---

[21] "Introduction: What Is Functionalism?" in Block, ed., *Readings in the Philosophy of Psychology* (Cambridge: Harvard, 1980), pp. 171-84.

> The local reductionist must grant that on his view there is nothing intrinsic that all pains have in common in virtue of which they are all pains (assuming that $N_h \vee N_r \vee N_m$ [that is, various neural realizations of pain] have nothing in common). But that is also precisely the consequence of the functionalist view. That, one might say, is the whole point of functionalism: the functionalist, especially one who believes in MR [multiple realizability], would not, and should not, look for something common to all pains over and above H (the heart of functionalism, one might say, is the belief that mental states have no "intrinsic essence").[22]

Kim implies that a functional specification does not provide a genuinely satisfactory answer to Block's question.

The reason he gives is this. On the nonreductive view, if *E* is a mental property and *B* is its neural or microphysical base, then realizers for *E* can be found in *B* (at the level of *B*). This position can also allow that nondisjunctive actual realizing properties might be found in *B* for individual species or structure types, as long as there is no property in *B* that is not wildly disjunctive that realizes *E* generally, that is, that realizes every possible instance of *E*. The nonreductive materialist claims that none of this entails a genuine reduction of *E* to properties in *B*. The standard strategy for preserving *E* as unreduced is indeed to retain *E* as a functional mental property. But Kim advances an objection to this functionalist move as a way of preserving nonreductive materialism about the mental:

> How should we counter this line of argument? I think it will be helpful to consider the causal picture, and ask: What are the *causal powers* of *this*

---

[22] "Multiple Realization and the Metaphysics of Reduction," p. 332, italics added.

> *instance of E*, namely [a system] *s*'s having *E* on this occasion? If *s* has *E* in virtue of *E*'s realizer *Q*, it is difficult to see how we could avoid saying this: the causal powers of this instance of *E* are exactly the causal powers of this instance of *Q*.[23]

Here Kim cites what we might call the causal inheritance principle for functional properties:

> (CIP - FP) If a functional property *E* is instantiated on a given occasion in virtue of one of its realizers, *Q*, being instantiated, then the causal powers of this instance of *E* are identical with the causal powers of this instance of *Q* (*ibid.*, p. 16).[24]

In Kim's view, the problem with the functionalist picture is that the causal powers of any instance of *E* will be causal powers in the physical base—they will not, at the token level, be irreducibly mental causal powers. Hence functionalism cannot preserve the view that there exist causal powers that are in the last analysis irreducibly mental, and it is thus incompatible with a robust nonreductive materialism about the mental. Furthermore, Kim points out that given the genuine multiple realizability of the property *E*, the causal powers of the realizers of *E* in *B* will exhibit significant causal and nomological diversity, and for this reason the causal powers of *E* will exhibit such diversity. Thus, in his estimation, *E* will be "unfit to figure in laws, and is thereby disqualified as a useful scientific property." He concludes that the functionalist model cannot protect *E* as a property with a role in scientific laws and explanations.[25]

I am sympathetic to this general line of argument against a functionalist nonreductive view of higher-level causal powers.[26] But I have argued[27] that there is available a nonfunctionalist account of these higher-level powers that nevertheless remains nonreductive. First of all, a line of thought with antifunctionalist implications can be found in the deepest criticism of behaviorism that Putnam advances in his classic essay, "Brains and Behavior."[28] There he argues that we should characterize mental states in a way that conforms to our characterization of kinds in the natural sciences. In the case of *polio*, for instance, we have found a biological explanation for the dispositional features of this disease (for example, its symptoms), and we identify the disease with the underlying biological properties that provide this explanation. By contrast, behaviorists identify mental states with dispositions to behave in certain ways given particular stimuli, and not with underlying mental properties that explain these dispositions. Putnam recommends that we abandon behaviorism in favor of a conception that would characterize mental states in accord with the biological example.

---

[23] "Making Sense of Emergence," p. 16, italics added.

[24] I deny (CIP-FP) for the following reason. Suppose *M* is a mental functional property realized at the neural level by *N*, and *N* is actually realized at the microphysical level by P1, but is multiply realizable at the microphysical level by P1 and P2. (CIP - FP) would have it that the causal powers of this instance of *M* are identical with the causal powers of *N*, and also with the causal powers of P1. By transitivity, the causal powers of *N* would be identical with the causal powers of P1, but by my argument above they will not be.

[25] "Making Sense of Emergence," pp. 17-18.

[26] In addition, one might note in this context a point made against functionalism by Paul Churchland, that the functionalist version of antireductionism can preserve theories that should be eliminated, such as alchemy and phlogiston theory, as easily as it can preserve the mental states of commonsense psychology—"Eliminative Materialism and the Propositional Attitudes," this JOURNAL, LXXVIII, 2 (February 1981). 67-90, at pp. 78-81. But, in a nonreductive materialist view according to which mental states are irreducibly mental intrinsic properties, this sort of worry will not arise.

Soon thereafter, however, Putnam developed and endorsed functionalism, and indeed he came to expect that functional properties would yield explanations for the dispositional features of mental states, much in the way that a viral infection provides an explanation for the symptoms of polio:

> My own view is that psychological predicates correspond to *functional* properties of human beings and other sentient beings. The presence of these properties *explains* the clustering of what some have called the 'symptoms' and 'criteria' of the various psychological states and conditions.[29]

This claim, however, amounts to the view that dispositions of types of mental states can be explained by dispositions of the same or other mental state types. As I have argued,[30] this would amount to what is at best a weak sort of explanation, and not an explanation in terms of mental causal powers that is adequate by Putnam's own scientific realist standards. What might it be for a type of dispositional property at a particular level to provide an explanation for another type of dispositional property at that level? To oversimplify, *pain* might be functionally characterized as follows:

---

[27] "Why a Scientific Realist Cannot Be a Functionalist," *Synthese*, LXXXVIII (1991): 341-58.

[28] In *Philosophical Papers, Volume 2*, pp. 325-41; first published in R. J. Butler, ed., *Analytical Philosophy: Second Series* (New York: Blackwell, 1965), pp. 1-19.

[29] Putnam, "Language and Reality," p. 278. Putnam endorsed functionalism in "The Nature of Mental States," in *Philosophical Papers, Volume 2*, pp. 429-40; first published as "Psychological Predicates," in W. H. Capitan and D. D. Merrill, eds., *Art, Mind, and Religion* (Pittsburgh: University Press, 1967), pp. 37-48.

[30] "Why a Scientific Realist Cannot Be a Functionalist," pp. 348-50.

> the state caused by pinpricks and pinches, that causes the thought "I should avoid those stimuli from now on," and given the belief that it's O.K. to express one's pain, causes winces and utterances of "ouch."

Now consider the proposal to explain a dispositional feature of *pain*, such as the tendency to cause winces, by means of the essential features of *pain*. Given functionalism, one would then be explaining *pain*'s tending to cause winces by a set of dispositions, and indeed, by a set that includes the very disposition to be explained as one of its components. Such an "explanation" is at best weak, for its crucial feature is a redundancy that undermines its strength.[31] In this respect, it differs markedly from explanations that are adequate by scientific realist standards.[32]

But the model inherited from other sciences, to which Putnam appeals in "Brains and Behavior," is best interpreted as explaining dispositional properties of kinds not simply by dispositional properties, but rather in part by properties intrinsic to those kinds—indeed, intrinsic properties proper to the same level as the kinds themselves. (A property is intrinsic to a kind just in case it is an intrinsic property of every possible instance of that kind.) In chemistry, the dispositional features of compounds are explained in part by their intrinsic molecular structure—chemical properties intrinsic to those kinds of compounds. In biology, polio symptoms, such as its contagiousness—a dispositional feature—are explained partly by an intrinsic biological property of that kind of disease, a particular viral infection. By analogy, the nonreductivist might consider the possibility that there are properties intrinsic to mental state types that play a part in explaining their dispositional features.

---

[31] This argument has the conclusion that the general fact that pain causes winces cannot adequately be explained by the general features of pain supposing that it is a functional kind. But this does not show that there is no context in which a particular wince might be adequately explained by pain, supposing that it is a functional kind.

[32] Block argues for a stronger antifunctionalist thesis, that functional properties cannot be causally efficacious in standard cases (cases in which no intelligent being recognizes them) and not only, as I have contended, that they cannot adequately explain dispositional features of kinds at the type level—"Can the Mind Change the World?" in *Meaning and Method* (New York: Cambridge, 1990), pp. 137-70. Functional properties, he argues, are second-order properties, and second-order properties cannot be causally relevant in standard cases. Examples help confirm this point. The property *dormitivity* is a classic example of second-order property, defined as follows: *x is dormitive* = *x* has some property that tends to cause sleep.

When a dormitive pill is slipped into your food, it is the chemical property of the pill, not dormitivity, that causes sleep. In addition, Block points out that the relation between a second-order property and an effect is logical, and not typically nomological. Consider the proposal: having some property that tends to cause sleep caused him to sleep. This claim would seem to specify a logical relation, and not (necessarily) a causal one. There might be a nomological relation in some such cases, but in most examples of this kind there seems to be no reason to posit one. Cf. Ruth Millikan, "Historical Kinds and the 'Special Sciences'," *Philosophical Studies*, XCV (1999): 45-65.

Richard Boyd[33] develops and defends a theory of kinds inspired by Putnam's claims. He notes that on various antirealist views, natural kinds are characterized in terms of observable features and dispositional properties. Boyd then argues that a position of this sort fails to account for successful inductions based on natural kinds—it does not allow for an explanation for the high degree of projectibility of these kinds. The remedy is to characterize natural kinds in terms of underlying causal powers that serve to explain their observable features and dispositional properties, and thereby to build grounds into the nature of these kinds for the success of inductions that appeal to them. Boyd has argued that in mature sciences, natural kinds are, in fact, typically characterized by such *explanatory essences*:

> Kinds characterized by "explanatory essences" are also kinds from the point of view of inductive generalization: indeed, in mature sciences, kinds which are explicitly characterized in terms of explanatory essences are the overwhelmingly typical cases of inductively natural kinds. Kinds natural from the point of view of successful induction need not always be explanatorily natural kinds, but they must correspond in relevant respects to the (perhaps unobservable) properties and mechanisms which causally determine the observable properties of the subjects of empirical generalizations (*ibid.*, p. 642).

Given the models we have for kinds in the natural sciences, it is reasonable to suppose that these essences would include properties intrinsic to the kinds.[34]

Now, I suspect that most functionalists maintain that the causal powers that have a role in explaining the dispositional features of mental states are nondispositional properties of their realization bases. For example, many suppose that nondispositional neural properties, which instantiate neural causal powers, would serve to explain why being pinched causes wincing behavior. But if these causal powers are all nonmental, a robust sort of nonreductive materialist account of the mental is precluded, for then none of the causal powers would be essentially mental themselves.

---

[33] "Scientific Realism and Naturalistic Epistemology," in *Proceedings of the Philosophy of Science Association*, II (1980): 613-62.

[34] See Boyd's "Kinds, Complexity, and Multiple Realization," in *Philosophical Studies*, xcv (1999): 67-98, for his view on whether mental kinds can fit this model.

By contrast, the nonreductivist might endorse intrinsic mental properties that instantiate specifically mental causal powers.[35] Such a view would be incompatible with functionalism. It need not deny that there exist functional mental properties, or, more generally, relational properties of mental states, but it would endorse nonfunctional mental properties that, by virtue of the causal powers they instantiate, play an important part in explaining dispositional features of mental state types.[36] What would such mental properties be like? First of all, despite the prevalence of functionalism, it is quite natural to suppose that the phenomenal content of a sensation type is intrinsic to it, and that even if belief contents are partially extrinsically individuated, it remains natural to suppose that the content of a belief type is at least partially intrinsic to it. Moreover, we readily assume that behavior is causally explainable by way of intrinsic features of mental states, and this is at least consistent with the claim that these mental states have intrinsic properties that instantiate mental causal powers.

But how might mental states of this description be multiply realizable? Consider the analogy of a ball-piston engine, the most recent version of the rotary internal combustion engine, which has a specific internal structural configuration.[37] Characteristic of this engine is its having parts with particular shapes and rigidities, and these parts must be arranged in a particular way. These features are manifestly not functional relations that such an engine stands in; rather, they constitute intrinsic characteristics of this type of engine. At the same time, these characteristics are multiply realizable. The parts of the engine can be made of material of different sorts—as long as the material can yield, for example, the required shapes and rigidities. The ball-piston engine, then, has a nonfunctionalist intrinsic structure that instantiates its causal powers, but it nevertheless admits of distinct realizations.[38]

---

[35] Robert Van Gulick's conception of "higher order patterns" that "have a degree of independence from their underlying physical realizations" also inspires the sort of view I am developing here; cf. "Who's in Charge Here? And Who's Doing All the Work?" in John Heil and Alfred Mele, eds., *Mental Causation* (New York: Oxford, 1993), pp. 233-56, at pp. 249-56.

[36] "Why a Scientific Realist Cannot Be a Functionalist," pp 350-51

[37] See the website: http://www.ballpistonengine.com.

[38] I suspect that the same point can be made with Putnam's peg-and-board example, in "Language and Reality," pp 295-98. If Philip Kitcher is right about the irreducibility of certain biological properties, it would appear that analogous claims would hold for them; "1953 and All That: A Tale of Two Sciences," *The Philosophical Review*, xcIII (1984): 335-73.

By analogy, it might be that the heterogeneous physical realizations of the dog's and the human's belief *that there is danger nearby* exhibit a structure of a single type that is intrinsic to this kind of mental state, a structure that instantiates the causal powers of this belief. This structure may be more abstract than any specific sort of neural structure, given that it can be realized in distinct sorts of neural systems.[39] Perhaps this same structure can be realized in a silicon-based electronic system, and such a system could then also have the belief. Suppose one built a silicon system that replicated the capacities of and interconnections among neurons in a human brain as much as is physically possible, and then excited the system to mimic as closely as possible what happens when a human being has this belief about danger. Is it not a serious empirical possibility that this silicon state would realize the same belief, and have a structure that, conceived at a certain level of abstraction, is similar enough to the structure of the ordinary neural system for both to count as instantiations of the same structure type? It would certainly seem far from likely that nothing of relevance would be alike in these systems other than relations to perceptual inputs, behavioral outputs, and other mental states. At very least, in this case and more generally, it makes little sense to retreat to mere functional resemblance prior to investigating whether the relevant similarities extend to intrinsic properties.

To what degree does the position I have been developing differ from Kim's view? The common ground includes the notion that, if mental states are causal powers, then they cannot be functionally defined. But in addition, Kim allows for neural structure-specific reduction. His idea is that there may, for instance, be neural structures common to more than one species to which some class of a mental state type can be reduced. I suggest that we might indeed identify a mental state type with a structure, but a structure more abstract than any specific neural structure, and one that can potentially be realized by a silicon-based system. Kim envisions the reducing structure to be neural, or physical at a lower level yet. My proposal is that there are structure types which cannot be classified as specifically neural, but which must rather be categorized as mental, and which would be intrinsic properties of mental states.

An important challenge to this proposal is that such structures may not exist, that in general, no significantly homogeneous structure types correlate with what are intuitively the tokens of mental state types. For instance, it may be that any higher-level structures that instantiate the belief *that there is danger nearby* differ on the order of the way in which a cat and an ordinary mousetrap differ as instantiations of the kind *mouse-catcher*.[40] This would constitute a serious challenge to a robust nonreductive materialist conception, but the view would not yet be decisively undermined. For it may be that these structure types, although they fail to correlate neatly with our ordinary mental state categories, are not specifically neural structure types either. In that case, one might take advantage of what room there is for altering the ordinary system for classifying mental states, at least for the purpose of scientific psychology, and identify the distinct structures with distinct mental state types. Scientific reclassifications relative to ordinary categories are, after all, not unusual. Still, it may also turn out that in general, the only significantly homogeneous structure types to be found are essentially neural structure types. In that case, it is hard to see how there could still be irreducibly mental causal powers, and in my view Kim would then be right. But that result does not appear very likely to me, given the thought that the structure of a state of a silicon system, conceived at some level of abstraction, could be similar enough to the structure of a neural realization of a mental state for both to count as features of the same kind.

---

[39] On abstractness of this sort, see Boyd, "Kinds, Complexity, and Multiple Realization," pp. 91-96.

It seems odd that realists about mental states have so firmly endorsed functionalism, a model for the nature of those states which is so closely tied to a general antirealist point of view. Positivist antirealists advocated an operationalist characterization of natural kinds, defining them in terms of the causes and effects of their instances. Logical behaviorism provides a good example of this practice. Scientific realism rejects this conception, however, for on this view the instances of natural kinds enter into causal relations without those kinds being defined by those relations. Although functionalist characterizations of mental states are more sophisticated than those of its behaviorist progenitor, functionalism nevertheless fits squarely within this antirealist tradition. My alternative proposal is not novel in spirit. It simply recommends for mental states what realists typically advocate for natural kinds generally.

## IV. MULTIPLE REALIZABILITY AND PROJECTIBILITY

It was once commonly supposed (by me, among others) that nonreductive views about the special sciences are grounded most fundamentally in the phenomenon of multiple realizability by way of a formal sort of argument. Kinds in the special sciences can be realized in different ways from the perspective of lower-level sciences, and thus an attempt to reduce higher-level kinds, laws, and explanations to those at a lower level will involve replacement by disjunctive properties—properties that are perhaps even wildly disjunctive in the sense that the disjuncts have at best little in common. Moreover, the disjunctions that these properties feature might even be open-ended or infinite. The received wisdom was that such disjunctive properties are not kinds, for the reason that statements of regularities involving such disjunctive properties fail to be laws, and perhaps most fundamentally, because "explanations" involving such disjunctive properties are not genuine explanations. This standard argument for nonreductive materialism appears to rely on a certain formal prescription for laws and explanations, that they cannot contain disjunctive properties, or at least not wildly disjunctive properties.[41]

---

[40] Cf. Boyd, "Kinds, Complexity, and Multiple Realization," pp. 71-72.

Kim argues, however, that a higher-level property is precisely as projectible as the disjunction that expresses its multiply realizable character at a more basic level, and thus a generalization involving such disjunctive properties is just as lawlike as the higher-level generalization that it was meant to reduce.[42] The reason is that a higher-level property is nomically equivalent to such a disjunctive property. Nomic equivalence might be defined in this way: properties *F* and *G* are nomically equivalent if they are coextensive in all possible worlds compatible with the laws of nature.[43] If Kim is right, then the formal argument does not appear to go through, for it relies on the possibility that generalizations involving a higher-level property be lawlike while those involving the corresponding disjunctive property are not. But furthermore, Kim contends that wildly disjunctive properties are not projectible, and hence higher-level properties that are nomically equivalent to such properties are not projectible either. As a result, such higher-level properties cannot figure into laws, and they are not genuinely scientific kinds.

The example of a disjunctive property Kim adduces to make his point is *jade*. 'Jade' is a category that comprises two mineralogical kinds, *jadeite* and *nephrite*, and hence *jade* is the same property as *jadeite or nephrite*. As a result, *jade* will not be projectible. For suppose that we are trying to confirm the generalization 'jade is green'. We might check many instances of jade and find that they are all green. But it could be that the entire sample consists of jadeite, and no nephrite. We must conclude that the generalization is not confirmed, and thus *jade* is not projectible.[44]

---

[41] Jerry Fodor, "Special Sciences," *Synthese*, XXVIII (1974): 97-115. Kornblith and I endorsed a version of this argument in "The Metaphysics of Irreducibility," pp. 126-28, against which William Jaworski has advanced an impressive counterargument—"Multiple Realizability, Explanation, and the Disjunctive Move," *Philosophical Studies*, CVIII (2002): 298-308, especially pp. 301-03.

[42] "Multiple Realization and the Metaphysics of Reduction," pp. 319-25.

[43] Block, "Anti-Reductionism Slaps Back," *Philosophical Perspectives*, XI (1997): 107-32, at p. 109.

To clarify his claim, Kim considers the objection that we can think of genuinely projectible kinds as disjunctive properties. *Emerald*, for example, can be thought of as *African emerald or non-African emerald*. But, he says, this possibility fails to undermine the projectibility of *emerald*—for example, it does not show that there is anything wrong with the lawlikeness of 'All emeralds are green'. But this analogy does not serve to reinstate the projectibility of *jade*, for, by contrast with "jadeite or nephrite,"

> the disjunction, "being an African emerald or non-African emerald," does not denote some heterogeneously disjunctive, nonnomic kind; it denotes a perfectly well-behaved nomic kind, that of being an emerald! There is nothing wrong with disjunctive predicates as such; the trouble arises when the kinds denoted by the disjoined predicates are heterogeneous, "wildly disjunctive," so that instances falling under them do not show the kind of "similarity," or unity, that we expect of instances falling under a single kind.[45]

But given this analysis, even *jade* might turn out to be a kind after all. As Block points out, all samples of jade share certain appearance properties, similarities that give rise to a certain degree of projectibility.[46] In Block's view, more generally, properties that are multiply realizable can yet be projectible with respect to "properties of channeled selection, learning, and design."[47] Because there are typically only a few ways in which entities of a particular higher-level type can be designed and produced, we can expect relatively broad similarities among these things which would render corresponding higher-level properties significantly projectible.[48]

---

[44] Fodor argues that Kim's example merely involves a sampling error—"Special Sciences: Still Autonomous after All These Years," *Philosophical Perspectives*, XI (1997): 149-63, at pp. 151-52.

[45] "Multiple Realization and the Metaphysics of Reduction," p. 321; cf. Lenny Clapp "Disjunctive Properties: Multiple Realizations," this JOURNAL, XCVIII, 3 (March 2001): 111-36, at pp. 120-21.

[46] "Anti-Reductionism Slaps Back," pp. 126-27. Louise Antony and Joseph Levine make a similar point in "Reduction with Autonomy," *Philosophical Perspectives*, XI (1997): 83-105, at pp. 90-91.

[47] "Anti-Reductionism Slaps Back," pp. 120-29.

The point I want to extract from this debate is that the heterogeneity of the possible realizations of a property is compatible with their having significant features in common, features that will undergird the projectibility of the property to some degree or other.[49] This point is consistent with Kim's claim that a higher-level property is precisely as projectible as the disjunctive property that comprises all of its possible realizations. One should not conclude from the heterogeneity of the possible realizations of a higher-level property that there is no feature that can sustain its projectibility—in fact, of both the higher-level property and of the disjunctive property that comprises all of its possible realizations. Indeed, the projectibility-sustaining feature of a kind could be a structure that is significantly homogeneous across its heterogeneous realizations, a structure that might instantiate a unitary causal power at the level of description of the kind. Note that disjunctive terms will typically fail to express or will at least mask any such homogeneous structural features and unitary causal powers to which they might correspond. In the case of the kind *ball-piston engine*, for example, a disjunctive term that details its possible realizations would fail to express or would at least mask the characteristic structural features on which the projectibility of this kind is based. By contrast, the term 'ball-piston engine' itself can serve to express these structural features without obscuring them.

But note that one cannot conclude merely from the fact that a property is projectible that it is an intrinsic structural feature that instantiates a unitary causal power at the level of description of the property. Functional properties, for example, may be projectible while they comprise neither intrinsic features nor unitary causal powers. *Being soluble* is projectible, yet although for any instance of this property there will be a causal power that will explain its projectibility, it does not itself instantiate a unitary causal power. The tie between projectibility and unitary causal powers is therefore looser than entailment. Whether a property is strongly or weakly projectible, it may turn out to be a functional property and have no intrinsic features that can instantiate a unitary causal power, let alone a unitary causal power at the level of description of the property.

---

[48] Cf. Antony and Levine, pp. 92-94.

[49] Cf. Clapp, pp. 123-32. He provides a fine account as to how a disjunctive predicate can indicate a nondisjuctive property on a causal powers notion of properties.

Consider an instructive example of Kornblith's. In 1869, the term 'neuraesthenia' was introduced to designate a nervous disease that results in severe fatigue—a characterization that is at least fairly functional. The term was soon established worldwide, but "like most descriptive terms, where basic organic or psychological understanding was lacking, it tended to be overinclusive and a receptacle for many diverse conditions."[50] But when cures for neuraesthenia were sought, it was found that different sorts of causes had to be treated. Several distinct sorts of underlying causes were discovered for the dispositional features of this purported natural kind. As a result, the term 'neuraesthenia' became obsolete by about 1930.[51]

What makes us think that *neuraesthenia* fails to instantiate a unitary causal power at the level of description of this property? First, there is the evidence that it is not projectible to a high degree. In addition, researchers discovered a disjunction of properties coextensive with *neuraesthenia* each of which is more strongly projectible. Explanations involving these properties effectively replaced those involving *neuraesthenia*. Moreover, the characterization of neuraesthenia was forced to remain fairly functional because no homogeneous underlying intrinsic features were discovered across its instances that could explain its dispositions and surface features. Accordingly, I would suggest that whether there is good evidence that mental states instantiate unitary and specifically mental causal powers depends on whether mental state types are projectible to a high degree, on the failure of a search for coextensive sets of properties that are more strongly projectible, and on whether intrinsic and specifically mental explanatory essences can be found. In short, whether there exists good evidence of this sort depends on whether there are powerful, resilient, and thoroughly realist psychological explanations in which mental state types play a part.

## V. THE SIGNIFICANCE OF MULTIPLE REALIZABILITY

If the multiple realizability of a mental property fails to support the claim that its realization base is at best only weakly projectible while the mental property itself is strongly projectible, does there remain a role for multiple realizability in the argument for nonreductive materialism? Let us first consider whether multiple realizability of mental state types might still sustain the nonreductive view, whereupon we will turn to the multiple realizability of mental tokens.

---

[50] L. Diamond, "Neuraesthenia," in Benjamin Wolman, ed., *International Encyclopaedia of Psychiatry, Psychology, Psychoanalysis, and Neurology, Volume VIII* (New York: Aesculapius, 1977), pp. 27-28.

[51] Kornblith, *Knowledge without Foundations: A Causal Theory*, unpublished Ph.D. dissertation, Department of Philosophy, Cornell University, Ithaca, New York, 1979, pp. 120-23; Chatel and Peele, "The Concept of Neuraesthenia," *International Journal of Psychiatry*, IX (1970): 36-49.

(i) *Types*. Lawrence Shapiro[52] is skeptical about any such significance for multiple realizability of mental state types:

> Take what appears to be a legitimate case of multiple realization.... Either the realizing kinds truly differ in their causally relevant properties, or they do not. If they do not, then we do not have a legitimate case of multiple realizability.... If the realizing kinds do genuinely differ in their causally relevant properties, then they are different kinds...and so we do not have a case in which a single kind has multiple realizations (*ibid.*, p. 647).

To illustrate the notion of a causally irrelevant property, Shapiro points to the color of a corkscrew. Corkscrews can be grey or black, for example, but the color of a corkscrew is causally irrelevant to its nature—in this case, to what it does. He argues on this basis that differences in color among corkscrews do not amount to a legitimate case of multiple realization.

Shapiro extends the claim of causal irrelevance to an example involving neural and silicon realizations of a mind. "If each neuron's contribution to psychological capacities is solely its transmission of an electrical signal, and if silicon chips contribute to psychological capacities in precisely the same way, then the silicon brain and the neural brain are not distinct realizations of a mind" (*ibid.*, p. 645). Here, he thinks, the sameness in contribution to psychological capacities "screens off" the difference between neurons and silicon chips, and makes it the case that they are not distinct realizations of these capacities. Legitimate cases of multiple realization of *E* would have to feature realizations of *E* that differ in their causally relevant properties, and realizations differ in their causally relevant properties only when they make distinct contributions to the nature of *E*.

But first of all, would it be inconsistent to claim that the identical contributions to psychological capacities made by neural and silicon systems do amount to a case of multiple realizability while denying this of the grey and the black corkscrews? No. The color of a corkscrew is causally irrelevant in a starker sense than the one Shapiro has in mind, for the color makes *no positive causal contribution whatever* to its nature—in this case, to what it does, never mind different colors not making *distinct causal contributions* to what it does. Shapiro's characterization extends causal irrelevance to pairs of realizers, each of which in fact makes a causal contribution to the nature of the thing, whenever each makes the same causal contribution. This characterization fails to count as different realizations pairs of distinct realizers each of which does, in fact, make a causal contribution to the nature of the thing that has it. So he says: "Steel and aluminum are *not* different realizations of the waiter's corkscrew because, relative to the properties that make them suitable for removing corks, they are identical" (*ibid.*, p. 644). But unlike color, being made of steel or of aluminum plausibly does make a causal contribution to what a corkscrew does—in this respect these properties are causally relevant in a way in which colors are not. Suppose that an effective corkscrew can only be made of steel and of aluminum, for only these materials have the right kind of rigidity, but that they make exactly the same contribution to what it does. The fact remains that these materials, as opposed to any others, have the right kind of rigidity. Accordingly, *making a causal contribution to the nature of the thing that has it* might be the notion of causal relevance that is pertinent to a condition on multiple realization. This alternative conception would license steel and aluminum but not distinct colors as multiple realizations of a corkscrew, and for silicon and neural systems to count as multiple realizations of psychological features.

---

[52] "Multiple Realizations," this JOURNAL, XCVII, 12 (December 2000): 635-54.

Why adopt this alternative conception of multiple realizability? Perhaps it is enough to point out that distinct realizers can make the same positive causal contribution to the nature of a thing, and this is just what we mean when we talk about multiple realizability. But in addition, the fact that the neural system and the silicon system make identical contributions to psychological capacities seems to force us to say that the features thus contributed are neither essentially neural nor essentially silicon-structural. Here there is significant work for multiple realizability to do: because some one type of thing can have realizations of distinct types *F* and *G*, it can be characterized neither as essentially *F* nor as essentially *G*. Shapiro is in no position to deny any of this. But given his conception, the realizations in this example will not really be multiple, and thus his conception fails to allow multiple realizability to do this work—which it can in fact do. For this reason, his conception of multiple realizability is best rejected in favor of the proposed alternative.

What then is the legitimate role for multiple realizability in supporting nonreductive materialism? The answer is implicit in the above discussion. Whether or not a property is multiply realizable can indicate the level at which it should be classified. Is the kind *corkscrew* a kind of steel thing? No, for it also has a possible aluminum realization. Is the kind *mind* a neural kind of thing? If mental states are also realizable in silicon, then no. Multiple realizability might then provide the key to precluding classification of mental states as essentially neural, or as essentially classified at some lower level yet.

Note that the realizability of mental states in both neural and silicon systems would not all by itself establish a robust nonreductive materialism. In addition, we would at least need evidence that mental states have intrinsic properties that might then instantiate essentially mental causal powers. For only then could we show from the evidence that mental states are realizable in non-neural systems that there indeed exist mental causal powers that are not essentially neural.

Now, I was surprised to discover several reductionist-leaning philosophers contending that physical types common to neural and silicon systems might well be discovered, but that this would lend *support* to reductionism. Oron Shagrir[53] adopts this line of argument, as do William Bechtel and Jennifer Mundale:

> A computer that could exemplify sufficiently similar behavior to biological organisms to justify the imputation of psychological states is likely to be very different from the ones we humans have designed to date, and the characterizations we would have to employ of its physical operations might turn out to be far more similar to those we use of brains than we currently expect. Such machines, for example, would likely have areas devoted to processing different sensory inputs and controlling motor outputs; potentially this might provide a basis for a common taxonomy of the physical processing states underlying psychological functioning (*op. cit.*, p. 204).

The advocate of a robust nonreductive materialism would welcome such a common taxonomy of these physical processing states, and so if the point that these advocates of "reductionism" are making is representative of their position, perhaps we can all agree! I would want to emphasize, however, that this common taxonomy will not be essentially neural, and so neural reductionism would be precluded. True, this taxonomy will be physical, but no more physical than a taxonomy for irreducibly neural states, which the nonreductive materialist also welcomes. The term 'physical' has a narrow and a broad sense: in the narrow sense, it picks out specifically anything over which theories in the science of physics quantifies; in the broad sense, it distinguishes anything wholly constituted from entities that are physical in the narrow sense. Nonreductive materialism is, after all, a kind of materialism, and hence does not countenance anything (in whatever realm it is claimed to hold true) that is not physical in the broad sense. If these reductionists are merely contending that for mental states shared by a brain and a computer we may find a common taxonomy that is physical in the broad sense, they are not proposing a hypothesis that is incompatible with nonreductive materialism.[54] Perhaps at this juncture these reductionists and some of their nonreductive opponents might find common ground.

---

[53] "Multiple Realization, Computation, and the Taxonomy of Psychological States," *Synthese*, CXIV (1998): 445-61, at pp. 454-55.

Now, Patricia and Paul Churchland[55] have argued that the multiple realizability of psychological states by type distinct neural states and by both neural and non-neural states does not undermine reductionism, for the reason that reductionism might be "domain-specific":

> ...visual experience may count as one thing in a mammal, and a slightly different thing in an octopus, and a substantially different thing in some possible metal-and-semiconductor android. But they will all count as visual experiences because they share some set of abstract features at a higher level of description. That neurobiology should prove capable of explaining all psychological phenomena in humans is not threatened by the possibility that some *other* theory, say, semiconductor electronics, should serve to explain psychological phenomena in *robots*. The two reductions would not conflict. They would complement each other (*ibid.*, p. 78).

If indeed visual experience in humans and in mammals had only functional characteristics in common, then the claim that they have distinct reductions would be plausible, for these distinct varieties of visual experience would not share causal powers. But if these varieties of visual experience share intrinsic structures, then they could share causal powers that are essentially neither neural nor electronic, but rather psychological or mental. Visual state types could then be mental state types that conform to the robust nonreductive materialist conception. A key motivation for the Churchlands' remarks is their sense that it is neuroscientific research that will reveal the nature of human psychology. But there is a natural way of understanding this motivation so that it is consistent with the nonreductivist picture I have been developing. If there exist irreducibly mental intrinsic structures, then a very likely avenue for discovering them would indeed be research in neuroscience. For even if such structures could also be realized in silicon-based electronic devices, it is highly plausible that their neural manifestations would first be discovered—precisely by way of neuroscientific research.

---

[54] Some might want to contend that these common features would not be psychological because they would have a physical-structural description, and anything that has this sort of description is *per se* not psychological. In this conception, anything that is psychological is describable only in familiarly mentalistic terms. But this view, advocated by some of the logical positivists, mistakenly ties the notion of the psychological too closely to a type of vocabulary. Moreover, the debate between nonreductive and reductive materialists should not be conflated with the debate between reductive materialists and those who hold that the mental is in no sense physical. (For an opposing view, see R. Endicott, "On Physical Multiple Realization," *Pacific Philosophical Quarterly*, LXX (1989): 212-24.) The claim that mental states are not identical with states that have a physical/structural description akin to descriptions of states in sciences such as biology and chemistry should not be regarded as an essential feature of nonreductive materialism.

[55] "Intertheoretic Reduction," in their *On the Contrary* (Cambridge: MIT, 1998), pp. 65-79. They cite temperature as an example of a property that has been domain-specifically reduced—their claim is that temperature is reduced differently in a gas, a solid, and in a vacuum. In my view, the differences among temperature in a gas, a solid, and a vacuum indicate that temperature is multiply realizable, but not that it has distinct reductions. For as Kornblith and I have argued, there is a unitary property that these cases of temperature have in common, with which temperature should be identified (*op. cit.*, pp. 138-39).

Nevertheless, various reductionists have, in my estimation, provided appreciable reason to believe that realizability in different kinds of neural systems alone need not advance the cause of robust nonreductive materialism. Suppose that what would seem to be a single mental state type were realizable only neurally, albeit in neural systems that differ, such as a human's, a dog's, and an octopus's. Imagine first that a single structural property was found that is intrinsic to this mental state type. This property might well count as neural and not as irreducibly mental, for as Bechtel and Mundale point out, within the realm of the neural itself there are possibilities for classification at different levels of abstractness or coarse-grainedness, and the property might well be characterizable as neural at some sufficiently high level of abstractness (*op. cit.*, pp. 201-04).[56] Then we would not have multiple realizability at the neural level after all. In addition, as Shagrir argues, if distinct neural mechanisms were found, then it might be that they correspond to distinct types of mental states, although it may initially have appeared that a single type of mental state was at issue (*op. cit.*, pp. 451-52).

(ii) *Tokens.* Even if types of mental causal powers are not identical with types of neural causal powers, it still could be that in the case of a normal human being, for example, every token mental power is identical with a neural causal power. (Earlier, I argued that token mental causal powers are not identical with token microphysical causal powers, but I reserved judgment on the token identity thesis for neural causal powers.) This is important, for if our mental causal powers were token identical with neural causal powers, a robust nonreductive materialism would be precluded, for then in the last analysis there would, in fact, be no irreducibly mental causal powers. In order to foreclose this sort of token identity, token mental causal powers would have to be multiply realizable in the right way.

---

[56] Cf. Kim, "Phenomenal Properties, Psychophysical Laws, and the Identity Theory," *Monist*, LVI (1972): 190-01, at p. 190; L. Mucciolo, "The Identity Thesis and Neurophysiology," *Noûs*, VIII (1974): 327-42.

The claim that token mental states and causal powers are identical with token neural states and causal powers is more resilient than is sometimes supposed. Indeed, stable tokens (given a certain level of complexity) often retain their identity over certain changes in their constitutions and configurations, and, significantly, they enjoy a certain resiliency in the production of their characteristic effects under these changes. So, for example, my decision to ring the doorbell can plausibly survive changes in its realizing microphysical state, and nature has likely endowed it with a resiliency for producing its characteristic effects under these small changes. But to establish that a token mental state $M$ is multiply realizable at the microphysical level does not suffice to show that it is not identical with the token neural state $N$ that constitutes it.[57] For $N$ might be similarly realiz-

---

[57] As a case in point, consider Yablo's argument for the primacy of the mental tokens over physical tokens as causes of actions in his "Mental Causation." He argues that when an event, such as *my ringing the doorbell*, has prima facie as causes both a token mental event—a decision—, and a token physical event—the physical realization of the decision—, it might well be that the mental event is *the* cause, albeit that both events are causally relevant. In his conception, which event is the cause depends on considerations of *proportionality*. Yablo's characterization of proportionality is complex, but in this case the relevant component is that the cause be *required* for the effect. In his definition of this notion, $x$s are causes, $y$s are effects, and event $x^-$ < event $x$ just in case $x^-$ is less specific than $x$; for example *Socrates' drinking the potion* is less specific than *Socrates' guzzling the potion*:

> $x$ is required for $y$ just in case for all $x^- < x$, if $x^-$ had occurred without $x$, then $y$ would not have occurred (*op. cit.*, p. 276).

In Yablo's estimation, the physical realization of the decision is more specific than the decision itself, for the reason that there are many distinct possible physical realizations of the decision. So the decision—the mental event—is the lesser of the two. Is it true that if the mental event had occurred without its specific physical realization, then my ringing of the doorbell would not have occurred?

> Of course the decision had a physical determination $p$; but most people would also say, and I agree again, that it would still have been succeeded by the ringing, if it had occurred in a different physical way, that is, if its physical determination had been not $p$ but some other physical event. And this is just to say that $p$ was not *required* for the effect (*op. cit.*, p. 278).

The relevant alternative physical event is the physical determination (realization) of the decision in the nearest world in which it is different from its actual physical determination. The problem is that Yablo's argument all by itself does not establish that the cause will not be, for example, neural rather than mental. Here, it is important to distinguish two distinct "physical" realizations of the decision, the neural and the microphysical. I think Yablo would be right to claim that, if the decision had occurred in a relevant alternative microphysical way, the effect would still have occurred. So the cause will not be the actual microphysical realization of able at the microphysical level, and thus its being multiply realizable in this way is consistent with its being identical with *M*.

Whether our mental states are token identical with neural states, even if they are not type identical with neural states, is not a thoroughly obvious matter. But perhaps it can be shown that token mental state *M*, which is in fact realized by token neural state *N*, could have been realized by a token neural state that is type distinct from *N* given the best neural classification. Given that *N* would not be identical to a token of such a distinct neural type, *M* would then not be identical *N*. Or maybe it can be established that *M* could have been realized by a token state that is not wholly neural but at least in part a silicon structure. Since *N* would not be identical to a realizing token that is not wholly neural, *M* would then also not be identical to *N*. This last scenario does not seem implausible. Suppose that at some time in the future we are capable of fitting brain-damaged patients with silicon-structural prostheses for the damaged parts of their brains. Now consider my desire to ring the doorbell, a token mental state, which, let us suppose, is wholly realized by a neural token. Would it have been the same mental token had it been realized in a neural and silicon structure rather than in a purely neural structure, holding all else fixed as much as possible? To my mind, it could well have been the same mental token, while at the same time the actual neural realizing token would be distinct from the neural and silicon realizing token. If all of this is indeed plausible, then it might well be that the mental token is distinct from its neural realizing token, and, more generally, that the mental/neural token identity thesis is false.

## VI. THE PROSPECTS FOR A ROBUST NONREDUCTIVE MATERIALISM

Kim and others have developed a number of very strong challenges to the nonreductive materialist position. In my estimation, the most daunting of these are the argument from explanatory exclusion and the contention that the functionalism that many nonreductive materialists espouse cannot accommodate irreducibly mental causal powers. I have attempted to answer these challenges, but one aspect of the antifunctionalist response bears highlighting. Common-sense

---

the decision. But this still leaves both the neural realization and the mental event itself as candidates for the cause. And the neural realization cannot be dismissed on the grounds that the decision would have been succeeded by the ringing if it had occurred in a different microphysical way, but only if it would have been succeeded by the ringing had it occurred in a different neural way. And whether it can be dismissed in this way is not immediately obvious. This worry about Yablo's argument can also be raised against Kripke's concerns about the token-identity thesis (*Naming and Necessity*, pp. 144-48), which Yablo cites (p. 269).

functionalist characterizations of mental states need not await the results of scientific investigation. Hence, if such functionalist characterizations did capture the nature of mental states, and nonreductive materialism could accommodate functionalism, then in an important respect nonreductive materialism need not await the results of scientific investigation for its validation. Kim is right, however, to claim that functionalism does not countenance mental causal powers, and for this reason functionalism is incompatible with a robust nonreductive materialism. As I have argued, mental state types would instead have to feature structural properties intrinsic to those types, and those structural properties must be distinct from any nonmental properties. But whether there exist structural properties of this sort is indeed a matter for empirical investigation, which is currently incomplete, and for this reason one's confidence that a robust nonreductive materialism is true might have to be moderated. If it turns out that there are no intrinsic properties of the right sort to be found, the nonreductive materialist will be forced toward reductionism. In the meantime, if my responses to the counterarguments are plausible, one might be confident that a robust nonreductive materialism about mental states remains a serious option.

DERK PEREBOOM

University of Vermont

# TRUST OUT OF DISTRUST[*]

Ever since people started reading Thomas Hobbes[1] with game theory-informed eyes, it became commonplace to see his state of nature as a prisoner's dilemma (PD-) structured situation. Starting out with his assumption of thoroughgoing egoism and given his description of the state of nature, Hobbes's conclusion is considered compelling: that human beings are destined to be locked into a state of mutual suspicion and distrust, of war of all against all. This state is a stable equilibrium, where distrust dominates trustful cooperation.

The interpretation of Hobbes's state of nature as a PD-structured situation has given rise to a puzzle: How can cooperation, which is based on trust, arise out of the distrust in which human beings seem to be irredeemably locked? After all, social life exists and depends on cooperation, and cooperation requires trust. To be sure, trust and cooperation are not identical, and neither of them is a necessary or a sufficient condition for the other. Yet there is significant overlap between the two. Much of what we regard as trustful behavior also falls under cooperative behavior. This is especially so when the relationships involved are thin rather than thick: when there are no special ties of kinship, friendship, a shared past, and the like among the people who are facing the choice of whether or not to cooperate trustfully with one another.[2] In a Hobbes-like state of nature, where relationships are thin and the options crude ("cooperate with others" or "fend for yourself"), cooperation more or less converges on trust. 'I trust you' reduces to my belief that you will do your part in the cooperative venture that serves the interests of us both, where the success of this venture depends on each of us doing our part.

But how do people mired in distrust and suspicion succeed in bootstrapping themselves onto promise keeping and trust? Hobbes's

---

[*] I wish to thank Robert Aumann, Russell Hardin, Dimitri Landa, Ariel Rubinstein, Frederick Schick, Cass Sunstein, and also the participants in the New York University Colloquium on Law, Philosophy, and Social Theory (fall 2001) and its conveners Ronald Dworkin, Thomas Nagel, and Jürgen Habermas. I am especially indebted to Piero La Mura and Gil Kalai for their generous help with the mathematical and game-theoretical aspects of this article, and to Avishai Margalit for his role in its shaping and sharpening throughout.

[1] *Leviathan*, C. B. MacPherson, ed. (New York. Penguin, 1982).

[2] For an analysis of trust in thick relationships, see, for example, Russell Hardin, "Distrust," *Boston University Law Review*, LXXXI, 3 (2001): 495-522; and my "Trust, Distrust and in between," in Hardin, ed., *Distrust* (New York: Sage, 2002, forthcoming).

0022-362X/02/9910/532–48

own solution, in the form and figure of the sovereign, is often regarded as somewhat of a deus ex machina and hence as not altogether satisfactory. The sovereign is external, and no satisfactory mechanism is or can be provided for his installment within the Hobbesian framework. And so the puzzle persists.

It is my aim here to establish the possibility of trust from within the Hobbesian framework. I shall show that distrust can be structured in two different ways: the first is commonly associated with Hobbes, and the second is sometimes associated with that other patron saint of social-contract theory, Jean-Jacques Rousseau. The two structures are here referred to as *hard* and *soft*, respectively, and they are both in essence compatible with Hobbes's stark assumptions about human nature. Hard distrust is PD-structured: the distrust strategy is dominant in it. Soft distrust is not PD-structured: the choice of trust in it is an equilibrium choice. In order to establish the possibility of trust, I do not need to claim or show that the state of nature is soft rather than hard, nor even that it is likelier to be so. All that is needed for trust to get off the ground is ambiguity, or uncertainty, as to which of the two basic situations of distrust obtains.

## I TWO SIMPLIFIED GAMES OF DISTRUST

What explains some people's tendency to consider distrust safer than trust? Why does some of the literature on trust take it almost for granted that "fairly generalized distrust might make sense in a way that generalized trust does not," and that suspicion and distrust are "inherently well grounded"?[3] There are, in principle, two alternative presumptions for dealing with cases of doubt about trust, one in favor of trust and one in favor of distrust. If one of them is to prevail as the default presumption, why is it often instinctively felt that the presumption of distrust should be the one?

Consider the following rough calculation of best and worst scenarios. First, the case of trust: acting on trust when trust is reciprocated can lead to successful cooperation and hence to mutual benefit and potentially to significant gain. Acting on trust when trust is not reciprocated and when one's partner is untrustworthy inevitably involves disappointment. It often involves worse: being betrayed or exploited. This may lead to serious damage. Consider next the case of distrust: acting on distrust when distrust is reciprocated leads roughly to whatever gains or benefits one is able to achieve on one's own. But what does acting on distrust lead to when distrust is not reciprocated (that is, when one's partner is trustworthy and trusting)?

---

[3] Hardin, p. 500; while these are quotes from Hardin's article, they do not express Hardin's view.

Here we may want to look at two different possibilities. One possibility involves a structure in which unilateral distrust, while entailing foregone opportunities, basically yields to the distrusting partner the same benefits as yielded by reciprocated distrust. On a natural interpretation of this type of case, the distruster is an individualist who "goes it alone," indifferent as to whether his partner chooses to trust or to distrust him. In the other possibility, the structure is such that unilateral distrust does benefit the distruster, at the expense of her trusting partner. In this type of case, the truster is being exploited by her partner. It is not two psychological types that are differentiated here, but rather two different structures of an interactive situation, as will now be shown.

*II.A. Soft distrust.* Here is a simple two-person payoff matrix to represent the structure of the first type of situation:

|  | T | DT |
|---|---|---|
| T | *4*;4 | *0*;2 |
| DT | *2*;0 | *2*;2 |

Figure 1: Soft distrust (lack of trust)

The two strategies open to each of the players are T for trust (or cooperation) and DT for distrust (or defection). If I as row-chooser (payoffs in italics) choose to trust you, I derive a payoff of 4 when you trust me, too, but as low as 0 when I am a lone truster. If I choose to distrust you, I derive a payoff of 2 regardless of whether you trust me or not. And the same goes for you, column-chooser.

This is a version of the stag-hunt game, which is famous in game theory as well as in philosophy[4]: "Separately we can catch rabbits and eat badly. Together we can catch stags and eat well. But if even one of us deserts the stag hunt to catch a rabbit, the stag will get away; so the other stag hunters will not eat unless they desert too" (*ibid.*, p. 7; cf. 95n.).[5] The credit for the original version goes to Rousseau,[6] who intended the story as a prototype of the social contract.

---

[4] See David Lewis, *Convention* (Cambridge: Harvard, 1969).

[5] For further discussion, see my *The Emergence of Norms* (New York: Oxford, 1977), pp. 121-24.

[6] *A Discourse on Inequality*, Maurice Cranston, trans. and intro. (New York: Penguin, 1984).

So the trust strategy (T) can lift you high or it can make you fall. It can be disappointing. Its alternative strategy (DT) is less risky: (DT) guarantees a payoff of 2, as compared with the risk of getting 0 for choosing (T). Distrust is here basically constant. It leaves you on some plateau that is insensitive to changes in your environment. At the same time, it is insensitive to the disappointments you may cause to others. There are no possibilities for you to exploit the other, nor are you in danger of exposing yourself to being exploited by the other. To the extent that 'safe' means hedging your bets, minimizing your potential losses, being risk-averse—distrust here seems indeed to be safer.[7] Both (T; T) and (DT; DT) are (Nash) equilibria. While the first equilibrium point is efficient in that it promises higher payoffs for both players, the second is the minimax choice.

But let us now generalize this situation and think of it not as a one-round encounter but rather as a repeated game with pure strategies. If the players start by playing it safe and adopting distrust, they may remain stuck with the suboptimal equilibrium in the future repetitions of the game. Suppose, however, that they succeed in coordinating on the better equilibrium, whether through communicating with each other or by responding to some commonly observed cues, or by any other way. Then they will both reap the higher fruit of trustful cooperation and neither of them will be tempted to deviate to distrust in the future repetitions of the game.

As was already noted, in this version of the two-person trust-distrust game, the truster is not exploited by the distruster. While the unilateral truster, whose trust is not reciprocated, suffers a loss (0), the distruster's payoff does not increase at her partner's expense (it remains 2 regardless of what the other player does). Strictly speaking, then, this game captures soft distrust; namely, it is a situation where one's unilateral employment of the DT strategy reflects what may be interpreted as mere lack of trust. Distrust here involves foregoing opportunities: while causing modest harm, it mostly protects against it.[8]

Let us generalize the interactive situation further: not just from a one-round to a repeated situation but also from two participants to a community of $n$ players. If we imagine this community to be repeatedly involved in situations of the soft-distrust kind, we may speculate that over time they will tend to develop "soft distrusting" dispositions. We may imagine them as a Wild West community of rugged individualists, honest folks who rely on no one and exploit no one. At the same time, however, there is no need to think of them as blind or averse to the possibility of joint ventures requiring trustful cooperation (like constructing a dam, say) and to their mutual benefits.

---

[7] Not only risk aversion but also loss aversion may play a role here. If you do not trust, you fail to get certain gains, as compared to your starting point; if you distrust, you protect yourself from losses, as compared to your starting point. For most people, losses are much "more bad" than gains are "good." So there might be a connection between a presumption of distrust and loss aversion.

[8] Compare Hardin, pp. 495-96

Note, moreover, that the payoff matrix of figure 1 and the story of the generalized soft distrust which goes with it may be given two different interpretations: a threshold one and an all-or-nothing one. (These are, in fact, well-known variants of the *n*-player stag hunt.[9]) In the first, in order to enjoy the higher payoff it is *not* necessary that all *n* participants choose the T-strategy: it is enough that the number, or proportion, of participants who choose T goes above a certain threshold. Think of the case, say, of daylight saving time, where those who do not switch their clocks are those who suffer the inconvenience, once a sufficient number have switched. In the second interpretation, in order to enjoy the higher payoff, it *is* necessary that all *n* participants choose the T-strategy. Here the interpretation of the matrix is as in the original story of the stag hunt mentioned earlier.

*II.B. Hard distrust.* Let us now consider the second way to conceptualize the two-person trust/distrust situation. This version captures distrust of the hard variety, where one's unilateral trust turns out to benefit the distruster at the expense of the truster.

| | T | DT |
|---|---|---|
| T | *4*;4 | *−2*;6 |
| DT | *6*;−2 | *2*;2 |

Figure 2: Hard distrust (betrayal of trust)

If I as row-chooser (payoffs in italics) choose to trust you (T), I derive a payoff of 4 when you trust me, too, and our cooperation succeeds. But I get as low as −2 when my trust is not reciprocated: my trust here may be seen as betrayed and exploited. If I alone choose to distrust you (DT), I in effect exploit you. My payoff is as high as 6, to your –2. When we both distrust each other, each of us must settle for 2, which is what we can achieve on our own. Once again, the game is symmetrical between row-chooser and column-chooser.

---

[9] See, for example, Hans Carlsson and Erik van Damme, "Equilibrium Selection in Stag Hunt Games," in Ken Binmore, Alan Kirman, and Piero Tani, eds., *Frontiers of Game Theory* (Cambridge: MIT, 1993), pp. 237-53.

It should come as no surprise that this is a PD-matrix. After all, lone defection in the canonical PD-situation is, for either prisoner, precisely what amounts to betrayal of the other prisoner's trust. Mutual trust is not an equilibrium point here: it is unstable. If no credible way can be devised for trust to be thrust upon the players, this jointly desired point remains all but inaccessible to them (in the one-round version of this game).

Except for the case where this game is indefinitely repeated, the distrust strategy (DT) is dominant.[10] It is hence, a fortiori, the safe strategy: trust has no chance here. It should be noted, however, that this statement holds so long as we focus on strict theory, and on agents whose behavior is restricted by the requirements of instrumental rationality. In practical PD-type situations, boundedly rational agents will find routes to cooperate and to justify cooperation.[11] Many experimental results, as well as ordinary experience, attest to this discrepancy between the predictions of (game) theory and actual behavior.[12]

In fact, more than risk aversion and loss aversion may be at work here. People are often also highly "betrayal averse" in the sense that they will tend to punish those who betray their trust. Indeed, the punishment for injury that comes from betrayal of trust will often be more severe than for an equivalent injury that involves no betrayal of trust.[13]

---

[10] More precisely: the DT strategy is the only strategy that survives the process of iterated elimination of dominated strategies. See D. Fudenberg and J. Tirole, "Perfect Bayesian Equilibrium and Sequential Equilibrium," *Journal of Economic Theory*, LIII (1991): 236-60.

[11] See, for example, Reinhard Selten, "The Chain Store Paradox," *Theory and Decision*, IX, 2 (1978): 127-59, where Selten promotes the distinction between *rational* (theory-driven) behavior and *reasonable* (practically oriented) behavior. This echoes the point made early on, by Duncan Luce and Howard Raiffa, in *Games and Decisions* (New York: Wiley, 1957), pp. 97-102.

[12] This statement should be qualified, as nowadays the theory attempts to make somewhat different predictions between strictly dominated strategies, where trust has really no chance, weakly dominated ones, where it depends, and iteratively dominated ones (like cooperation in the finitely repeated PD), where it also depends. See J. Y. Halpern, "Substantive Rationality and Backward Induction" (2000), in ewp-game/0004008 (consult http://econwpa.wustl.edu/eprints/game/papers/0004/0004008.abs) where both Robert J. Aumann's and Robert Stalnaker's approaches to the epistemic conditions behind backward induction are compared and discussed.

If a distrust situation of the soft kind is pictured as conducive to a Wild West community of individualists, we may picture the distrust situation of the hard kind as conducive to a community of classical Western gunmen. A John Wayne type is forever suspicious, forever on the alert to be the first to draw his gun and to take advantage of any moment of weakness on the part of his mates, as well as to engage in self-defense against them.

## III. THE STATE OF NATURE

Two games or models of archetypal situations of distrust have been presented: soft and hard, or if you will, mild and harsh. Not every situation where the question of trust versus distrust comes up is PD-structured, and, as we saw, in those cases that fall under the soft category, mutual distrust is not a dominant strategy. In fact, in this class of cases, mutual trust not only leads to a jointly beneficial outcome but it is a strict equilibrium and, as such, it is accessible to the participants. Still, when people think of paradigmatic cases of trust versus distrust, it is often PD-structured situations that they have in mind, namely, cases of the hard variety. To many people, it would seem almost axiomatic that a general distrust strategy dominates a general trust strategy inasmuch as being a truster is often considered much worse than simply taking a chance: it is actually taken to *mean* being a sucker by exposing oneself to exploitation.

The general outlook of hard distrust may well derive from the powerful hold that Hobbes's grim picture has over us, the picture of the state of nature as a state of suspicion of all in all and of a war of all against all. A Hobbesian will tend to interpret many social interactions, whether micro or macro, as one-round PD-structured games. And it is this general outlook that motivates a general presumption in favor of distrust.

But even Hobbes himself, in presenting what he calls the "precept, or general rule of Reason," distinguishes between two situations. He says, first: "Every man ought to endeavor Peace, as farre as he has hope of obtaining it"; and then he continues: "and when he cannot obtain it, that he may seek, and use, all helps, and advantages of Warre." The first part of this precept contains what is for Hobbes the Fundamental Law of Nature: *to seek peace and follow it.* The second part sums up what he refers to as the Right of Nature: *by all means we can, to defend our selves* (*op. cit.*, chapter XIV). There is nothing far-fetched or strained, as far as I can see, in interpreting the first part of the precept as applying to situations of soft or mild distrust, and the second as applying to situations of hard or harsh distrust.

---

[13] For more on this, see Jonathan J. Koehler and Andrew D. Gershoff, "Betrayal and Aversion" (unpublished manuscript, 2000). I owe this point to Cass Sunstein, whose example is our attitude to injury caused by our own car's airbag.

True, Hobbes did not believe that one's endeavoring peace in the hope of obtaining it would get one very far. A close reading of the relevant passages reveals how deeply convinced he was that we are doomed constantly to seek and use the advantages of war to defend ourselves. But still, the important point is that he did seem to recognize the possibility that the state of nature be construed in terms of soft distrust as well as hard distrust. And, as I shall now proceed to show, this possibility is all we need in order to establish the possibility of trust.[14]

In principle, if we are able to diagnose our situation as belonging to either the hard or the soft category, we are in familiar territory. But what if we cannot tell? More intriguingly, what if we do not know to which category the state of nature belongs—if we do not know, that is, which of the two games nature has put us in? The possibility of such an ambiguity is my leading idea in establishing the possibility of trust.

## IV. THE GAME OF NATURE: SOFT OR HARD?

Suppose that we are playing a distrust game. We know that the game we play is either hard or soft distrust, but we do not know which of the two it is. (A reminder: going clockwise from top left, the payoff matrix of soft is (4;4), (0;2), (2,2) (2;0) and of hard it is (4;4) (−2;6) (2;2)

---

[14] There is a question whether the adequate model for the Hobbesian state of nature is that of *compulsory* interactions or that of *optional* ones. As introduced by Philip Kitcher—"The Evolution of Human Altruism," this JOURNAL, XC, 10 (October 1993): 497-516—an optional game is a situation in which a player is allowed two degrees of freedom: the possibility of signaling which partners she is willing to play with, and the possibility of opting out of the game altogether. Now, it may well be that the evolution of altruism (or of trust) works better, along the lines suggested by Kitcher, in the context of repeated optional PD-games than of repeated compulsory ones, where players have neither degree of freedom. The account I shall offer here is not evolutionary, and it is meant to apply to the pessimistic framework of a Hobbesian state of nature with nonoptional interactions. A fortiori, then, it will apply to the optional framework, where discriminating partners—prepared to trust anyone who has not played DT in a past interaction with them—would be expected to do better. (Note, however, that on my interpretation of the two distrust games, the choice of DT amounts to "going it alone," and hence to opting out of the game altogether. This signals the existence of the intermediate category of nonoptional yet *partially compulsory* games, with just one degree of freedom: you have no choice about your partner/opponent, but you can opt out. This category seems relevant in both evolutionary and nonevolutionary accounts of the emergence of trust.)

(6;−2); hard is a PD, soft is not; the strategies are T for trust and DT for distrust; the strategy (T;T) is an equilibrium in soft but not in hard.) How are we to think about our situation? How should we play? And, crucially, might trust prevail?

If both of us play T, each of us gets 4 regardless of which game we are playing, and if both of us play DT, we get 2 each; in either case with a symmetrical choice of strategy, we shall still not know which game we are playing. Only if we play asymmetrically, that is, if one of us plays T and the other plays DT, will we know for sure which game we are playing. Let us look at a story that is meant to make vivid the possibility that two people might find themselves not knowing whether the situation they are in is, in fact, one of hard or soft distrust.

Two survivors of an air crash find themselves on the desert side of an island. Separately, they can survive by farming, each having their own well of water. Before they crashed, however, they were able to see that across the desert there is a river, the other side of which looks opulent and green. But crossing the river requires a bridge, and it takes two to build a bridge. Let us assume that the payoff for each islander from staying put and farming is 2, and from crossing to the other side of the river, where the living is (or promises to be) easy, is 4. Now suppose they agreed that, if they were to survive the crash, they would meet at the river and join forces to cross it. Al sets out on his journey across the desert, trusting that Bill will do the same. But what if Bill does not reciprocate?

We are to imagine two situations. In the first, Al realizes after a while that he has been stood up and he returns to farm his land. He pays a price for having crossed the desert in vain and for having lost farming time, so his payoff is 0, while Bill's remains 2. (Of course, the situation is symmetrical between them: had Bill been the lone truster, his payoff would be 0 to Al's 2.) This corresponds to the soft-distrust game. In the second situation, it turns out that there is a subterraneous connection between Al's and Bill's wells. This means that when Al sets on his journey and stops farming, Bill has double the amount of water at his disposal. Moreover, while Bill's crops become better and more bountiful, Al's farm dries up. Here, then, Bill not only benefits from Al's lone trust, but his benefit is at Al's expense: his payoff goes up to 6, while Al's decreases to −2. This situation corresponds to the hard-distrust game.

Note that the different way distrust plays out in the two situations derives from the objective circumstances in which the islanders find themselves, and not from their being differently motivated. When Bill stands Al up, he certainly proves himself untrustworthy. But it is not as if he can be described as "exploitative" when he does it in the second case as compared with his being more "benign" in the first. Rather, it so *happens* that in the second case Bill gains at Al's expense, whereas in the first case he does not. And, of course, as long as Al's and Bill's choice of strategy is symmetrical—that is, they either both stay put and farm, or they both set out for the river—they will not know whether they are in a soft or hard situation.

We now go back to the general case of a distrust game, where it is known that the game to be played is either hard or soft, but it is not known which of the two it is. Suppose that the probability $p$ of the occurrence of the game of soft distrust is known. When this is the case, the two games of hard and soft can be superimposed on one another to yield one compound game: we can calculate the payoffs accordingly, check for equilibrium points, and decide on the best strategy.

| | T | DT |
|---|---|---|
| T | 4;4 | $p \times 0 + (1-p) \times (-2)$;<br>$p \times 2 + (1-p) \times 6$ |
| DT | $p \times 2 + (1-p) \times 6$;<br>$p \times 0 + (1-p) \times (-2)$ | 2;2 |

Figure 3: The compound game of hard and soft distrust

With these particular payoffs, if $p \geq 0.5$, that is, if chances are half (or more) that the probability that the game we are facing is soft, then the trust strategy T remains an equilibrium strategy. (The minimal value of $p$ that sustains T as an equilibrium strategy will vary with varying payoffs.)

*IV.A. Infinitely repeated game.* Suppose now that this game situation recurs: we are to play this game repeatedly. That is, we go from a one-round game to a *super game,* and we are further to suppose that the repetitions we are talking about are infinite (or at least with no fixed horizon in view). If the game we play is soft, then T is an equilibrium strategy for each round, regardless of whether the game is played once or repeatedly. Repetition does make a difference, however, if the game we play is hard, which is a PD-type game. Here, as is well known, even though in the one-round game DT is the dominant strategy, in the repeated game the trustful strategy T may be an equilibrium outcome.[15] Both players come to realize that there is less to gain by optimizing (that is, choosing DT) in the short run than to lose in the long run. Consider the strategy, known as *tit for tat*, that says: "I will play T in the first round and then do whatever my partner/opponent did in the previous round." This strategy, if adopted by both of us, is (Nash) equilibrium in the infinitely repeated game.

So it turns out that in the infinitely repeated case, T is an equilibrium outcome regardless of whether the game we play is soft or hard. And this means that *trust* will remain an equilibrium outcome in our compound super game, where we do not know which of the two games we are playing. Note that so long as we adopt the above strategy and stick to it without mistakes or glitches, we shall not know which game it is that we are playing. Nor will we have any incentive to want to know.

*IV.B. Finitely repeated game.* Things are different, however, when there is a finite and definite horizon. At least, things are different with regard to the hard game: with regard to soft, trust is an equilibrium strategy regardless of whether the game is played once, or infinitely many times, or finitely many times. But with hard PD-structured games, as soon as it is common knowledge between us that we shall play the game no more than, say, 1000 or 100 times, the entire cooperative scheme might "unravel from the back."

Consider: if we get to round 100 and it is common knowledge between us that it is the last, then "there is no longer-term loss to weigh against the short-term gain," and both of us will "defect" to DT. But then when we get to round 99, we both know that we shall play noncooperatively in round 100, and the same consideration applies to the current round as well; and so on, by so-called backward induction, to the very first round. "To get cooperation, there must always be a future substantial enough to outweigh immediate considerations."[16] The conclusion seems to be that in the finitely repeated hard game, there is no escaping distrust.[17] As was noted earlier, this conclusion is not necessarily borne out in practice: people do not betray each other as often and as harshly as the game-theoretic analysis predicts or prescribes. Cooperation in iterated PD, when the number of iterations is large, may even be the norm. And in any case, "unless the whole fixed-number iteration is extremely pristine, one cannot get the deduction going from the *n*th to the first."[18]

---

[15] This, roughly, is what is known as the *folk theorem*. It holds in this formulation as long as both players are patient enough, that is, when their "discount factor" is close enough to 1. For more details and technical niceties, see, for example, David M. Kreps, *A Course in Microeconomic Theory* (New York: Harvester Wheatsheaf, 1990), pp. 503-15; and Martin J. Osborne and Ariel Rubinstein, *A Course in Game Theory* (Cambridge: MIT, 1994), pp. 143-46. Note: it turns out that T is also an equilibrium outcome in yet another version of the infinite PD-game. Instead of imagining one pair playing indefinitely, we imagine a sequence of PD-games being played by pairs of people in a society with an infinite number of members. Everyone gets to play against everyone else, and each pair plays only once. The trustful cooperative solution applies here as it does in the case of PD being repeatedly played between the same two partners. The equilibrium strategy is: I play T unless somebody plays DT with me, in which case I switch to DT against my next partner and stick to it. The threat of "punishment" here is indirect, as I shall never play against you again, but you will play many times against people who will have played with me, so you will eventually be harmed. See, for example, Michihiro Kandori, "Social Norms and Community Enforcement," *Review of Economic Studies*, LIX (1992): 63-80.

But now we come to consider the case where the game that is finitely repeated is our *compound* game. If we do not know which of the two distrust games we are playing, and we are to play the game finitely many times, what strategic recommendations apply? Specifically, might trust be an equilibrium choice here?

Trust may well prevail in the simple situation where there is common belief between us that chances are that we are playing soft (say, that $p > 0.5$). But now suppose that this is not the case. Specifically, we are interested in the case where we are both "pessimistic," in the sense that, in Hobbes's spirit, we both think chances are that the game we are facing is hard rather than soft. Various situations might be considered; I shall offer two.

*(i) Chivalry*. Let us consider first the following setup: one of our two games (or states of nature) is picked by nature, and played $N$ (say, 1000) times. The probability that nature picked soft is a small epsilon ($\epsilon$), the probability that the game to be played is hard is $1 - \epsilon$. The outcomes and payoffs in each round are commonly observed by both players, he and she. Consider the following strategy: both players play T up until $k$ steps before the end of play. Now $k$ steps before the end,

---

[16] This quote, as well as the previous quotes in this paragraph, are from Kreps, p. 514.

[17] More precisely, the conclusion is that trustful cooperation cannot be a subgame-perfect equilibrium. But, in fact, the stronger result holds as well, that trustful cooperation cannot even be a Nash equilibrium, as can be shown by a different argument; see Osborne and Rubinstein Note, also, that in the case of a society where every pair plays only once, if the number of players is finite the backward induction will once again push them back to DT.

[18] Hardin, *Collective Action* (Baltimore: Johns Hopkins, 1982), p. 149. For his dissenting view of the backward induction argument, see Hardin's discussion there, at pp. 145-50. See also Philip Pettit and Robert Sugden, "The Backward Induction Paradox," this JOURNAL, LXXXVI, 4 (April 1989). 169-82; and Cristina Bicchieri, "Self-Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge," *Erkenntnis*, XXX (1989): 69-85.

that is, in round $N - k$, he continues to play T but she switches to DT. Since this is an asymmetrical strategy and the payoffs are commonly observed, both players will now know which game it is that they are playing. If it turns out that they are playing hard, the strategy calls for them both to play DT in the remaining $k - 1$ steps. If it turns out they are playing soft, the strategy is that they both play T in the remaining $k - 1$ steps. (And if either player deviates at any point from the prescribed behavior, then the two players switch immediately to DT and stick to it until the end of the game.[19])

This strategy is (Nash) equilibrium.[20] The optimal $k$ is calculable: it depends only on the probability $\epsilon$ (that the game is soft) and on the given payoffs, but not on the horizon of play.[21] Thus, for the particular payoffs we have been considering, and with a "pessimistic" probability $\epsilon$ of 0.1 that the game is in fact soft, it turns out that it is optimal for both players to play T up until something like ten rounds before the end. In other words, even if the players are "pessimistic" and believe with high probability that they are, in fact, facing hard, then when $N$ is sufficiently large the mere possibility of soft may produce the desired bootstrapping which will block the backward induction and "pull" toward trust, for most of the duration of play. What makes a better equilibrium than DT possible here is the fact that knowledge is possible: at any stage it is possible for the players to find out what game they are playing (and also, finding out at some stage promises the players better payoffs than playing DT throughout). At the same time, interestingly, the Nash equilibrium that yields the best payoffs for both players is the one that postpones the moment of truth as far as possible.

Note that, although the setup we consider is symmetric between the two players, the possibility of knowledge, and hence the Nash equilibria, are essentially predicated on an asymmetric step: he continues with T, she deviates to DT. (There is, to be sure, at least one symmetric equilibrium point, namely, the least efficient one where they both play DT throughout.) Of course, the roles of the players can be reversed, or assigned to a commonly observed flipped coin, but this is still an asymmetric solution. This circumstance may be taken to make a case for the importance of certain social conventions, like chivalry (he sacrifices; she deviates) or deference to old age, and so on which may signal the direction of the asymmetry and thus facilitate a solution.

---

[19] This requires the tiny caveat that, if the deviating partner is she, in that she neglects to switch to DT when she is supposed to, then she is to switch to DT at the next step.

[20] This is, in fact, not one strategy but a family of strategies—for each choice of $k$. These strategies are Nash equilibria, provided $k$ is not too small (that is, if it is not too close to the end of play). The smallest $k$ for which this is a Nash equilibrium yields the highest payoffs overall for the players.

[21] See Osborne and Rubinstein, p. 156.

*(ii) "Perturbation": one-sided information.* Let us consider now a setup where knowledge is not possible: the payoffs in each round are not accessible to the players.

Once again, one of our two games (or states of nature) is picked by nature, and played $N$ times; soft is played with a small probability epsilon ($\epsilon$). There is a certain probability $p$ (or rather $(p - \tau)$, as explained below) that both players are informed of the game they are playing, and a certain probability $(1 - p)$ that both players are not informed. There is, however, also a small probability theta ($\tau$) that only one of them is informed. (We are to suppose that the uninformed remain uninformed until the end.)

Suppose first that he is uninformed. Now, if he observes that she chooses T in the first round, he might infer that she probably knows which game is being played and that she believes he knows as well. He might further infer that she must know that they are playing soft, and that this is why she is offering cooperation by choosing trust. But then he should also choose T from that point on (up until close to the end of play). Suppose next that he is informed. If his information is that the game they are playing is soft and he observes that she chooses trust, then the previous reasoning applies and he should choose trust as well. Consider, finally, the case when he is informed, and the information is that the game they are playing is hard. In this setup, if he observes that she chooses T in the first round, he will infer that she is probably uninformed and that, if he were to reciprocate with T, she might infer that he knows that they are playing soft and she will therefore continue to cooperate. Now, the entire chain of reasoning in this passage can, of course, be replicated for her by starting out supposing that she is uninformed and that she observes that he chooses trust in the first round.

So it is, in fact, in the interest of the uninformed to mimic the informed—or to mimic anyway on the belief that the other might be informed. And it is also in the interest of the informed to choose trust even when he or she is informed that they are playing hard, given that it is common knowledge between them that it is in the interest of the other to mimic. This will eventually "pull" toward coordination on mutual trust (almost to the end) even among the uninformed and even when the game picked by nature is in fact hard.[22] Note that, if it is common knowledge between them that they are both uninformed, they will just look at the expected payoffs. Given our assumption that $\epsilon$ is small, these look pretty much like the payoffs for hard, and so the trustful equilibrium will be destroyed and both players will distrust (DT). The point is that the "pull" toward coordination on trust will occur even when the players are in fact uninformed, provided that there is some uncertainty in the mind of each, however small, as to whether the opponent is informed or not.[23]

There is a theoretical result in the game-theoretical literature[24] that amounts, roughly, to the following. In the finitely repeated PD, a one-in-a-thousand chance that one's opponent is irrationally benevolent (in the sense that she is determined to stick to the strategy of tit-for-tat) completely changes the theoretical prediction that there will be an unstoppable unraveling from the back toward noncooperation from round 1. That is, when a tiny "perturbation" (or a "trembling hand") is introduced in the form of incomplete information about the rationality of player 1, then player 2 can know without inconsistency what player 1 will do because, with small probability, player 1—being irrationally benevolent—will permit this to be known. And "from this small wedge, we are able to get them both to play [the cooperative strategy] happily until near the end of the game."[25] Moreover, even if player 2 is not "irrational" in this fashion, the "rational" thing for player 1 to do is to mimic this sort of behavior, to keep cooperation alive.

---

[22] Note: there is always a Nash equilibrium in which neither the informed nor the uninformed ever trust the opponent. But the point here is that there cannot be an equilibrium in which the informed cooperate by choosing trust whenever possible and the uninformed never choose trust. If the informed cooperate, then it is also in the interest of the uninformed to mimic the informed, and this breaks the proposed equilibrium. Therefore, any other equilibrium besides the one of full distrust must involve mimicry, which "pulls" both informed and uninformed toward trust.

[23] The two examples of finitely repeated games here discussed are special cases of a general result of Benoit-Krishna: J.-P. Benoit and V. Krishna, "Nash Equilibria of Finitely Repeated Games," *International Journal of Game Theory*, XVI (1987): 197-204. See also the related discussion in Osborne and Rubinstein, pp. 155-60.

[24] Kreps, P. Milgrom, J. Roberts, and R. Wilson, "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma," *Journal of Economic Theory*, XXVII (1982): 245-52.

[25] Kreps, p. 541. Indeed, as it turns out, even minimal violations of the hypothesis of common knowledge are sufficient to justify the emergence of cooperation in the repeated PD. See Abraham Neyman, "Cooperation in Repeated Games When the Number of Stages Is Not Commonly Known," *Econometrica*, LXVI, 1 (1999): 45-65, and also Aumann, "Irrationality in Game Theory," in P Dasgupta, D. Gale, O. Hart, and E. Maskin, eds., *Economic Analysis of Markets and Games. Essays in Honor of Frank Hahn* (Cambridge: MIT, 1992), pp. 214-27 (reprinted in Aumann, *Collected Papers*, Volume 1 (Cambridge: MIT, 2000), pp. 621-34).

The distrust case just discussed, where there is a tiny element of uncertainty about asymmetric information, is structurally similar to this case of the finitely repeated PD-game, where rational cooperation is proven possible given a tiny perturbation of irrationality. The crucial difference, however, between the two is that, in our distrust case, the perturbation has nothing to do with irrationality.[26] In our distrust case, it enters through the narrow crack of the possibility that a different game is being played and that there is asymmetrical information between the players: that only one of them is informed which of the two games is being played. As we saw this possibility, even if tiny, is capable of sustaining an equilibrium strategy of trust almost to the end of play.[27]

## V. THE POSSIBILITY OF TRUST

In a society where there is war of all against all and people expect the worst of each other, people may come to anticipate that the types of interactions in which they will find themselves are PD-structured situations, and will act accordingly. In such a society, people will lock themselves forever in mutual suspicion and distrust. From this kind of distrusting animal, no trusting, promise-fulfilling creatures can hope to emerge. The pattern of behavior they adopt is self-reinforcing. In such a society, the possibility of being exposed to trust behavior and its benefits, and thereby to learning trust behavior, is effectively blocked.

---

[26] Pettit and Sugden show that the backward induction need not unravel from the back in the finitely repeated PD-game for a different reason than the trembling-hand perturbation. Indeed, they argue a stronger point, which is that "rational players are necessarily not in a position to run the backward induction argument" (*op. cit.*, p. 182), since (1) any act of cooperation by either player would cause the breakdown of an assumption that is required for them to run the induction—namely, their common knowledge/belief in each other's rationality, and (2) "there are beliefs which a player might rationally hold which would make it rational for him to cooperate initially" (*op. cit.*, p. 179). For a similar argument see Bicchieri. But this solution, while not relying on a "perturbation," still relies on irrationality (in the form of a breakdown in the common knowledge, or belief, in the rationality of the players).

[27] In the literature, the effect of such tiny perturbations of incomplete information is discussed also in connection with the other well-known games that exhibit a finite-horizon paradox: the centipede game and the chain-store game. It is also tied with the phenomenon of *reputation*: I may find it beneficial to sometimes act "irrationally" in order to build up a reputation, if being thought irrational might later be beneficial to me through the influence this has on how others play. See Kreps, pp. 536-43.

This pessimistic scenario can be avoided without resorting to Hobbes's solution, deus ex machina-style, of the external Sovereign. Nor need the assumptions about human nature be made more charitable. We may, without changing too much in Hobbes's initial assumptions, stick to his unflattering picture of human beings as basically egotistical creatures, and yet see a way for these creatures to have the resources for arriving at mutual trust. All they need is to realize that an alternative setup for distrust exists: to recognize that not all distrust situations are PD-structured. With this, there will emerge the recognition that, depending on the type of situation, "distrust" is ambiguous between a harsh, exploitative sense and a milder sense. The notion of egoism, too, will be seen to be open to an "isolationist," defensive interpretation which underlies a situation of soft distrust, in addition to an offensive interpretation, which underlies situations of hard distrust.

But the main point is that in order for trust to get going, no assumption need be made that, as a matter of fact, the state of nature is *not* Hobbes's PD-structured one. All that is required is that the possibility be acknowledged that it *may not* be so structured. The mere ambiguity between hard and soft distrust is all that is needed for trust to emerge.

EDNA ULLMANN-MARGALIT

Center for Rationality and Interactive Decision Theory/Hebrew University of Jerusalem

# NEW BOOKS

ABELA, PAUL: *Kant's Empirical Realism.* New York: Oxford, 2002. viii, 303 p. Cloth $55.00.

ADLER, JONATHAN E.: *Belief's Own Ethics.* Cambridge: MIT, 2002. xv, 357 p. Cloth $40.00.

ACHTENBERG, DEBORAH: *Cognition of Values in Aristotle's Ethics: Promise of Enrichment, Threat of Destruction.* Albany: SUNY, 2002. xiii, 218 p. Cloth $62.50, paper $20.95.

ALLINSON, ROBERT ELLIOT: *Space, Time, and the Ethical Foundations.* Burlington, VT: Ashgate, 2002. 203 p. Cloth $69.95.

ANDERSON, CHARLES W.: *A Deeper Freedom: Liberal Democracy as an Everyday Morality.* Madison: Wisconsin UP, 2002. xi, 208 p. Cloth $27.95.

BAILEY, ALAN: *Sextus Empiricus and Pyrrhonean Scepticism.* New York: Oxford, 2002. xv, 302 p. Cloth $40.00.

BASINGER, DAVID: *Religious Diversity: A Philosophical Assessment.* Burlington, VT: Ashgate, 2002. vii, 123 p. Paper $24.95.

BELL, RICHARD H.: *Understanding African Philosophy: A Cross-Cultural Approach to Classical and Contemporary Issues.* New York: Routledge, 2002. xviii, 189 p. Cloth $85.00, paper $21.95.

BONJOUR, LAURENCE: *Epistemology: Classic Problems and Contemporary Responses.* Lanham, MD: Rowman and Littlefield, 2002. viii, 291 p. Cloth $75.00, paper $23.95.

BOOLOS, GEORGE S., JOHN P. BURGESS, and RICHARD C. JEFFREY: *Computability and Logic,* Fourth Edition. New York: Cambridge, 2002. xi, 356 p. Cloth $75.00, paper $28.00.

BRAIDOTTI, ROSI: *Metamorphoses: Towards a Materialist Theory of Becoming.* Malden, MA: Blackwell, 2002. x, 317 p. Cloth $59.95, paper $24.95.

BROUGHTON, JANET: *Descartes's Method of Doubt.* Princeton: University Press, 2002. xv, 217 p. Cloth $26.95.

BUZAGLO, MEIR: *Solomon Maimon: Monism, Skepticism, and Mathematics.* Pittsburgh: University Press, 2002. xiii, 168 p. Cloth $32.50; *The Logic of Concept Expansion.* New York: Cambridge, 2002. xi, 182 p. Cloth $55.00.

CAIN, M.J.: *Fodor: Language, Mind and Philosophy.* Malden, MA: Blackwell, 2002. 240 p. Cloth $59.95, paper $22.95.

CAVANAGH, MATT: *Against Equality of Opportunity.* New York: Oxford, 2002. viii, 223 p. Cloth $35.00.

CHIMISSO, CRISTINA: *Gaston Bachelard: Critic of Science and the Imagination.* New York: Routledge, 2001. xi, 285 p. Cloth $100.00.

CLARK, MICHAEL: *Paradoxes from A to Z.* New York: Routledge, 2002. xi, 218 p. Paper $15.95.

COCKS, JOAN: *Passion and Paradox: Intellectuals Confront the National Question.* Princeton: University Press, 2002. xi, 220 p. Cloth $49.50, paper $16.95.

CUNNINGHAM, FRANK: *Theories of Democracy.* New York: Routledge, 2002. vii, 248 p. Cloth $75.00, paper $21.95.

DAN-COHEN, MEIR: *Harmful Thoughts: Essays on Law, Self, and Morality.* Princeton: University Press, 2002. x, 309 p. Cloth $65.00, paper $19.95.

DEUTSCH, ELIOT: *Persons and Valuable Worlds: A Global Philosophy.* Lanham, MD: Rowman and Littlefield, 2001. Cloth $80.00, paper $34.95.

DUDDY, THOMAS: *A History of Irish Thought.* New York: Routledge, 2002. xviii, 362 p. Paper $19.95.

EWIN, R. E.: *Reasons and the Fear of Death.* Lanham, MD: Rowman and Littlefield, 2002. vii, 167 p. Cloth $70.00, paper $24.95.

GALEOTTI, ANNA ELISABETTA: *Toleration as Recognition.* New York: Cambridge, 2002. viii, 242 p. Cloth $60.00.

GOVIER, TRUDY: *Forgiveness and Revenge.* New York: Routledge, 2002. x, 205 p. Cloth $80.00.

GREEN, MICHAEL STEVEN: *Nietzsche and the Transcendental Tradition.* Urbana: Illinois UP, 2002. xii, 199 p. Cloth $29.95.

FALZON, CHRISTOPHER: *Philosophy Goes to the Movies: An Introduction to Philosophy.* New York: Routledge, 2002. ix, 230 p. Cloth $75.00, paper $19.95.

FELDMAN, STEVEN P.: *Memory as a Moral Decision: The Role of Ethics in Organizational Culture.* New Brunswick, NJ: Transaction, 2002. xiii, 249 p. Cloth $39.95.

FISCHER, PHILIP: *The Vehement Passions.* Princeton: University Press, 2002. x, 268 p. Cloth $26.95.

FLANAGAN, OWEN: *The Problem of the Soul: Two Visions of Mind and How to Reconcile Them.* New York: Basic, 2002. xvi, 364 p. Cloth $27.50.

FLORIDI, LUCIANO: *Sextus Empiricus: The Transmission and Recovery of Pyrrhonism.* New York: Oxford, 2002. xvi, 150 p. Cloth $45.00.

FULLER, STEVE: *Social Epistemology,* Second Edition. Bloomington: Indiana UP, 2002. xxxi, 314 p. Cloth $49.95, paper $21.95.

FUMERTON, RICHARD: *Realism and the Correspondence Theory of Truth.* Lanham, MD: Rowman and Littlefield, 2002. xiii, 149 p. Cloth $60.00.

GALSTON, WILLIAM A.: *Liberal Pluralism: The Implications of Value Pluralism for Political Theory and Practice.* New York: Cambridge, 2002. ix, 137 p. Cloth $55.00, paper $19.00.

GAUKROGER, STEPHEN: *Descartes' System of Natural Philosophy.* New York: Cambridge, 2002. viii, 258 p. Cloth $60.00, paper $22.00.

GEANEY, JANE: *On the Epistemology of the Senses in Early Chinese Thought.* Honolulu: Hawaii UP, 2002. viii, 267 p. Paper $20.00.

GENSLER, HARRY J.: *Introduction to Logic.* New York: Routledge, 2002. ix, 399 p. Cloth $80.00, paper $24.95.

GILLOCH, GRAEME: *Walter Benjamin: Critical Constellations.* Malden, MA: Blackwell, 2002. xi, 304 p. Paper $22.95.

GOMEZ-LOBO, ALFONSO: *Morality and the Human Goods: An Introduction to Natural Law Ethics.* Washington, DC: Georgetown, 2002. xvii, 142 p. Paper $17.95.

GRAHAM, KEITH: *Practical Reasoning in a Social World: How We Act Together.* New York: Cambridge, 2002. xi, 202 p. Cloth $55.00.

HACKING, IAN: *Historical Ontology*. Cambridge: Harvard, 2002. vii, 279 p. Cloth $39.95.

HALLEN, BARRY: *A Short History of African Philosophy*. Bloomington: Indiana UP, 2002. 133 p. Cloth $35.00, paper $14.95.

HAMMER, ESPEN: *Stanley Cavell: Skepticism, Subjectivity, and the Ordinary*. Malden, MA: Blackwell, 2002. xiv, 199 p. Cloth $59.95, paper $22.95..

HAWLEY, KATHERINE: *How Things Persist*. New York: Oxford, 2001. ix, 221 p. Cloth $45.00.

HETHERINGTON, STEPHEN: *Good Knowledge, Bad Knowledge: On Two Dogmas of Epistemology*. New York: Oxford, 2001. x, 209 p. Cloth $45.00.

HINDE, ROBERT A.: *Why Good Is Good: The Sources of Morality*. New York: Routledge, 2002. xiv, 241 p. Cloth $80.00, paper $25.95.

HOFFMAN, JOSHUA, and GARY S. ROSENKRANTZ: *The Divine Attributes*. New York: Blackwell, 2002. viii, 204 p. Cloth $59.95.

HURLEY, S.L.: *Consciousness in Action*. Cambridge: Harvard, 2002. xii, 506 p. Paper $25.00.

HYMAN, ARTHUR: *Eschatological Themes in Medieval Jewish Philosophy*. Milwaukee: Marquette, 2002. 136 p. Cloth $15.00.

IERODIAKONOU, KATERINA: *Byzantine Philosophy and Its Ancient Sources*. New York, Oxford, 2002. vii, 309 p. Cloth $55.00.

JOSEPHSON, PETER: *The Great Art of Government: Locke's Use of Consent*. Lawrence: Kansas UP, 2002. xi, 366 p. Cloth $45.00.

JOYCE, RICHARD: *The Myth of Morality*. New York: Cambridge, 2001. xiii, 249 p. Cloth $60.00.

KAPLOW, LOUIS, and STEVEN SHAVELL: *Fairness versus Welfare*. Cambridge: Harvard, 2002. xxii, 544 p. Cloth $45.00.

KASULIS, THOMAS P.: *Intimacy or Integrity: Philosophy and Cultural Difference*. Honolulu: Hawaii UP, 2002. xii, 183 p. Paper $14.95.

KIRKMAN, ROBERT: *Skeptical Environmentalism: The Limits of Philosophy and Science*. Bloomington: Indiana UP, 2002. x, 211 p. Cloth $39.95, paper $19.95.

KLEMENT, KEVIN C.: *Frege and the Logic of Sense and Reference*. New York: Routledge, 2002. xiii, 260 p. Cloth $70.00.

KRAKOW, IRVING: *Why the Mind-Body Problem Cannot be Solved: Some Final Conclusions in the Philosophy of Mind!* Lanham, MD: UP of America, 2002. vii, 112 p. Paper $26.00.

KUKLICK, BRUCE: *A History of Philosophy in America, 1720-2000*. New York: Oxford, 2001. xiii, 326 p. Cloth $30.00.

LADYMAN, JAMES: *Understanding Philosophy of Science*. New York: Routledge, 2002. xiii, 290 p. Paper $75.00, cloth $21.95.

LANGE, MARC: *An Introduction to the Philosophy of Physics: Locality, Fields, Energy, and Mass*. Malden, MA: Blackwell, 2002. xvii, 320 p. Paper $25.50.

LAWLOR, LEONARD: *Derrida and Husserl: The Basic Problem of Phenomenology*. Bloomington: Indiana UP, 2002. xii, 286 p. Cloth $49.95, paper $19.95.

LEVY, DAVID J.: *The Integrity of Thinking*. Columbia: Missouri UP, 2002. ix, 152 p. Cloth $34.95.

LITCH, MARY M.: *Philosophy through Film.* New York: Routledge, 2002. vii, 242 p. Paper $22.95.

LONG, A.A.: *Epictetus: A Stoic and Socratic Guide to Life.* New York: Oxford, 2002. xiv, 310 p. Cloth $29.95.

MACKEY, LOUIS: *An Ancient Quarrel Continued: The Troubled Marriage of Philosophy and Literature.* Lanham, MD: UP of America, 2002. viii, 283 p. Paper $43.00.

MATTHEWS, ERIC: *The Philosophy of Merleau-Ponty.* Ithaca, NY: McGill-Queen's vi, 186 p. Cloth $75.00, paper $27.95.

MCGUINNESS, BRIAN: *Approaches to Wittgenstein: Collected Papers.* New York: Routledge, 2002. xv, 299 p. Cloth $80.00.

MCLEOD, CAROLYN: *Self-Trust and Reproductive Autonomy.* Cambridge, MIT, 2002. xiii, 199 p. Cloth $29.95.

MEDINA, JOSÉ: *The Unity of Wittgenstein's Philosophy. Necessity, Intelligibility, and Normativity.* Albany: SUNY, 2002. xii, 233 p. Cloth $62.50, paper $20.95.

MERCER, CHRISTIA: *Leibniz's Metaphysics. Its Origins and Development.* New York: Cambridge, 2001. xii, 528 p. Cloth $80.00.

MOLDOVEANU, MIHNEA C., and NITIN NOHRIA: *Master Passions. Emotion, Narrative, and the Development of Culture.* Cambridge, MIT, 2002. xx, 247 p. Cloth $29.95.

MOODY-ADAMS, MICHELE M.: *Fieldwork in Familiar Places: Morality, Culture, and Philosophy.* Cambridge: Harvard, 2002. x, 260 p. Paper $19.95.

MORISON, BENJAMIN: *On Location: Aristotle's Concept of Place.* New York: Oxford, 2002. viii, 194 p. Cloth $45.00.

## NOTES AND NEWS

Michael Kelly, Managing Editor of the *Journal of Philosophy* since 1987, has accepted the position of Executive Director of the American Philosophical Association. His tenure with the *Journal* has been marked by innovation and imagination, and he will be greatly missed by those who have worked with him. The Trustees and Editors of the *Journal*, grateful for his unflagging devotion to its excellence, wish the very best for him in his new position.

**Outstanding Scholarship from Cambridge**

**Kierkegaard and the Treachery of Love**
*Amy Laura Hall*
Cambridge Studies in Religion and Critical Thought 9
0-521-80913-4 Hardback $58 00
0-521-89311-9 Paperback $21.00

**The German Aesthetic Tradition**
*Kai Hammermeister*
0-521-78065-9 Hardback $60.00
0-521-78554-5 Paperback $22.00

**Off the Beaten Track**
*Martin Heidegger*
0-521-80114-1 Hardback $60.00
0-521-80507-4 Paperback $20 00

**Searching for Sustainability**
Interdisciplinary Essays in the Philosophy of Conservation Biology
*Bryan G. Norton*
Cambridge Studies in Philosophy and Biology
0-521-80990-8 Hardback $85.00
0-521-00778-X Paperback $30.00

**Adam Smith's Marketplace of Life**
*James R. Otteson*
0-521-81625-4 Hardback $70.00
0-521-01656-8 Paperback $26.00

**Kierkegaard, Religion and the Nineteenth-Century Crisis of Culture**
*George Pattison*
0-521-81170-8 Hardback $65.00
0-521-01042-X Paperback $24.00

**Radical Cartesianism**
The French Reception of Descartes
*Tad M. Schmaltz*
0-521-81134-1 Hardback $65.00

**Moral Philosophy from Montaigne to Kant**
First Edition
*J.B. Schneewind, Editor*
0-521-80259-8 Hardback $25.00
0-521-00304-0 Paperback $25.00

**Adorno's Positive Dialectic**
*Yvonne Sherratt*
0-521-81393-X Hardback $60.00

**The Cambridge Companion to Duns Scotus**
*Thomas Williams, Editor*
Cambridge Companions to Philosophy
0-521-63563-2 Paperback $23.00

**Paradox and Paraconsistency**
Conflict Resolution in the Abstract Sciences
*John Woods*
0-521-81094-9 Hardback $70.00
0-521-00934-0 Paperback $26.00

*Available in bookstores or from*
CAMBRIDGE
UNIVERSITY PRESS
800-872-7423
us.cambridge.org/philosophy

*New Philosophy Titles from . . .*

HB Humanity Books

**THE CULTURE OF EXTINCTION**
Toward a Philosophy of Deep Ecology
**Frederic L. Bender**

Global warming, air and water pollution, ozone-layer depletion, species extinction—these are all the results of what philosopher Frederic L. Bender calls our "culture of extinction." In this probing analysis Bender criticizes many of industrial society's basic assumptions and reveals the underlying connections between mechanistic science and Earth's devastation under capitalism. He also points out that major modernist philosophies, from Bacon and Descartes to 20th-century empiricism, are in the service of the culture of extinction.

455 pp • ISBN 1-59102-055-7 • HC $65

**HEGEL'S *PHENOMENOLOGY OF SPIRIT***
New Critical Essays
**Edited by Alfred Denker and Michael Vater**

This volume of essays offers an interpretation of the spirit of Hegel's *Phenomenology* as well as a concise reading of the main text. It also discusses the historical and philosophical background of Hegel's main work and takes note of its reception. Since the essays were written by philosophers from different countries—both established Hegel scholars and promising young researchers—the collection presents the reader with an international overview of recent Hegel research.

285 pp • ISBN 1-59102-056-5 • PB $28

***. . . and***

PB Prometheus Books

**THE INFLUENCE OF NATURAL RELIGION ON THE TEMPORAL HAPPINESS OF MANKIND**
**Jeremy Bentham**
*Introduction by Delos McKown*

GREAT BOOKS IN PHILOSOPHY SERIES

Bentham criticizes the major tenets of belief in a supreme being and its alleged benefits to humanity. He then catalogues the many ways in which natural religion harms both individuals and society as a whole. This little-known work was considered so controversial in its time that it was first published under the pseudonym of "Philip Beauchamp."

152 pp • ISBN 1-59102-033-6 • PB $12

**For exam copies contact editorial@prometheusbooks.com**
**Call toll free (800) 421-0351 • Fax (716) 691-0137 • www.prometheusbooks.com**

NEW REQUIRED READING FROM OXFORD

## PATHWAYS TO KNOWLEDGE

PRIVATE AND PUBLIC

ALVIN GOODMAN

Examines the notion of private and public knowledge, arguing for the epistemic legitimacy of private and introspective methods of gaining knowledge, yet acknowledging the equal importance of social and public mechanisms in the quest for truth Throughout, Goldman addresses this opposition but proposes a rigorous framework that continues to resolve such contradictions, making this exciting text one of the most important contributions to the theory of knowledge in recent years.

240 pp $35 00 cloth

## REFERENCE AND CONSCIOUSNESS

JOHN CAMPBELL

Investigates how consciousness of the world explains our ability to think about the world, how our ability to think about objects we can see depends on our capacity for conscious visual attention to those things. Campbell illuminates classical problems about thought, reference, and experience by looking at the underlying psychological mechanisms on which conscious attention depends

240 pp $19 95 paperback

*New In Paperback!*

## UNDERSTANDING HUMAN KNOWLEDGE

PHILOSOPHICAL ESSAYS

BARRY STROUD

Presents the best of Stroud's essays in human knowledge Seeks to identify the question that philosophical theories of knowledge are meant to answer, and the role skepticism plays in making sense of that question This volume is essential reading for anyone whose work touches on human knowledge

272 pp $19 95

## NOT PASSION'S SLAVE

EMOTION AND CHOICE

ROBERT SOLOMON

Reopens the classic debate about the role of sentimentality in art and literature, defends the often-neglected role of compassion in deliberations of justice, in management theory, and in education, the role of love as a touchstone of ethical behavior and demonstrates how emotions give meaning to and play an essential role in our lives

304 pp $35 00 cloth

*New In Paperback!*

## WHAT'S WITHIN?

NATIVISM RECONSIDERED

FIONA COWIE

"An excellent contribution to the often murky literature on nativism The exposition is clear and accurate, and the style is reader friendly Cowie draws the right distinctions, and her arguments are judicious and penetrating "
—Peter Alward, *Minds and Machines*

352 pp $21 95

## ENGAGING REASON

ON THE THEORY OF VALUE AND ACTION

JOSEPH RAZ

A penetrating exploration of the interdependence of value, reason, and the will The essays illuminate a wide range of questions concerning fundamental aspects of human thought and action

344 pp $19 95 paperback

*New In Paperback!*

## A THEORY OF ART

KAROL BERGER

"Berger's *A Theory of Art* is a *tour de force* of breadth, comprehension, and coverage Its argumentative style is eminently lucid, accessible, and honest "
—Lydia Goehr, Professor of Philosophy, Columbia University

304 pp $19 95

OXFORD

## Dostoevsky the Thinker

JAMES P. SCANLAN

"*Dostoevsky the Thinker* offers a fresh and persuasive view of Dostoevsky's thought, which it treats with deserved respect and critical distance, neglecting neither Dostoevsky's coherence and seriousness, nor his 'intellectual untidiness.' Organized around an elegant set of categories, its argument is clear, accessible, and well-grounded."—William Todd, Harvard University. 1 illustration. $29.95

## What Ought I to Do?

### *Morality in Kant and Levinas*

CATHERINE CHALIER

Translated by Jane Marie Todd

"Whereas most introductions to Levinas's thought take the reader by way of the history of twentieth century European philosophy, Catherine Chalier here succeeds in laying out Levinas's ethical philosophy by reference to Kant. [This book] succeeds in making Levinasian ethics accessible to the broader philosophical community."—Robert Bernasconi, The University of Memphis. $45.00 cloth, $17.95 paper

## The Analytic Imaginary

MARGUERITE LA CAZE

Marguerite La Caze explores the idea of the imaginary of analytic philosophy. Noting the marked tendency of analytic philosophy to be unselfconscious about the use of figurative language and the levels at which it works, La Caze shows how analytic images can work to define the parameters of debates and exclude differing approaches, including feminist ones. $32.50

## Naturalism Defeated?

### *Essays on Plantinga's Evolutionary Argument against Naturalism*

Edited by JAMES BEILBY

Almost a decade ago, Alvin Plantinga articulated his bold and controversial evolutionary argument against naturalism. This intriguing line of argument raises issues of importance to epistemologists and to philosophers of mind, of religion, and of science. In this, the first book to address the ongoing debate, Plantinga presents his influential thesis and responds to critiques by distinguished philosophers from a variety of subfields. $45.00 cloth, $19.95 paper

## Retrieving Experience

### *Subjectivity and Recognition in Feminist Politics*

SONIA KRUKS

"This splendid book may well become a landmark in the development of feminist theory. Drawing on the work of both postmodern and earlier existentialist writers, Kruks offers new solutions to the problems that have recently impeded theoretical and practical attempts to improve womens' situations."—Hazel E. Barnes, University of Colorado at Boulder. $36.50 cloth, $17.95 paper

## Rethinking Rape

ANN J. CAHILL

"*Rethinking Rape* is an extremely interesting, well researched, and accessibly written book on a significant and timely topic. Ann Cahill's focus on rape as seen through recent continental feminist theories is original and much needed in the literature." —Susan J. Brison, Dartmouth College. $39.95 cloth, $18.95 paper

Sage House 512 East State Street, Ithaca, New York 14850

### Damaged Identities, Narrative Repair

HILDE LINDEMANN NELSON

"Thoughtful, persuasively argued and witty, this important book analyzes powerful assaults on group identities, then proposes strategies of resistance and repair. A sustained, original, political development of narrative approaches to ethics and moral psychology."—Sara Ruddick, author of *Maternal Thinking: Toward a Politics of Peace*.
$43.50 cloth, $17.95 paper

### Political Forgiveness

P. E. DIGESER

"*Political Forgiveness* is the first book to develop a conception of and rationale for forgiveness as a strictly political rather than personal or economic response to a wrongdoing. P.E. Digeser has contributed an original framework for understanding an important topic."—Martha L. Minow, author of *Between Vengeance and Forgiveness*. $39.95

### The Ethics of Transracial Adoption

HAWLEY FOGG-DAVIS

"Arguments over the rules that should govern transracial adoptions ignite great passions. [This book] adds much-needed rigor to those debates. [It] displays deep philosophical reflection, an admirable willingness to accord respect to opposing perspectives, a keen eye for nuance, and lively, accessible writing that will provide an excellent basis for discussion in a wide range of disciplines."
—Randall Kennedy, Harvard Law School. $25.00

### Religious Schools v. Children's Rights

JAMES G. DWYER

"James Dwyer offers a fearless—and, I think, convincing—critique of parental rights in regard to the education of children. It is sure to generate controversy and, more importantly, to provide deep illumination about the way we conceptualize (and treat) children in our society."—Sanford Levinson, University of Texas Law School.
$15.95 paper

### Against Liberalism

JOHN KEKES

**An Alternate Selection of the Eagle Book Club**

"A comprehensive, pointed, and fair critique of the . . . assumptions that lie behind liberal politics, [this book] demonstrates why the old liberal idea of giving people the moral autonomy to choose and live out good lives must, because of its self-contradictions, fail."
—*Review of Metaphysics*.
$17.95 paper

### A Case for Conservatism

JOHN KEKES

"If one considers the moral suppositions of conservatism, there isn't anyone in academic life better prepared to illuminate these ideas than John Kekes. In [this book] he makes his argument with perspicacity, logical exegesis, and compelling argumentation."
—Herbert London. $17.95 paper

**Cornell University Press**

At bookstores, or call (607) 277-2211 • www.cornellpress.cornell.edu

# crítica

**REVISTA HISPANOAMERICANA DE FILOSOFÍA**

Vol. 34 / No 100 / Abril 2002 / ISSN 0011-1503

*Artículos*

AMÁN ROSALES RODRÍGUEZ, *El determinismo tecnológico y la "dialéctica de la historia"*

GERT HELGESSON, *What Is Implicit?*

IGNACIO ÁVILA CAÑAMARES, *El nuevo enigma de la inducción y los términos de clase natural*

LEE MCINTYRE, *Supervenience and Explanatory Exclusion*

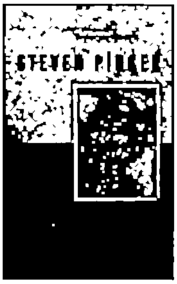
*Notas bibliográficas*

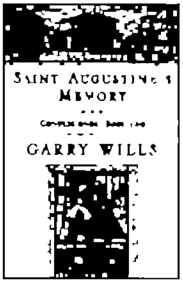
MARÍA LILIANA LUKAC DE STIER, *El fundamento antropológico de la filosofía política y moral en Thomas Hobbes* [Faviola Rivera Castro]

*Publicaciones recientes*

---

CRÍTICA, *Revista Hispanoamericana de Filosofía* is published in April, August, and December by the Instituto de Investigaciones Filosóficas, Universidad Nacional Autónoma de México. All correspondence should be addressed to CRÍTICA, Apartado 70-447, Coyoacán, 04510, D.F. México. e-mail: critica@filosoficas unam.mx





Søren Kierkegaard

### A LITERARY REVIEW

*Translated with an Introduction and Notes by Alastair Hannay.* The only English translation available in paperback of Kierkegaard's incisive essay on religion, nihilism, and post-modernism.

Penguin Classics original

116 pp 0-14-044801-2 $13.00

Steven Pinker

### THE BLANK SLATE

**The Modern Denial of Human Nature**

"A remarkable book—stimulating, fearless, immensely learned, and a pleasure to read."—Orlando Patterson, Harvard University.

Viking 528 pp 0-670-03151-8 $27.95

Saint Augustine

### SAINT AUGUSTINE'S MEMORY

**Confessiones, Book Two**

*Translated by Garry Wills.* The pivotal volume of the Christian philosopher's work, translated and interpreted by Pulitzer Prize winner and acclaimed Augustine biographer Garry Wills.

Viking 240 pp 0-670-03127-5 $23.95

*Available November 2002*

Voltaire

### MICROMÉGAS

**and Other Short Fictions**

*Newly Translated by Theo Cuffe. Introduction and Notes by Haydn Mason.* Features an introduction that places the stories in the context of Voltaire's writing and the moral, philosophical, and social questions they pose.

Penguin Classics original

192 pp 0-14-044686-9 $10.00

Elio Frattaroli, M.D.

### HEALING THE SOUL IN THE AGE OF BRAIN

**Why Medication Isn't Enough**

"Elio Frattaroli illuminates hugely important truths about our innate human complexity in this fascinating excursion through psychology, neuroscience, literature and quantum physics."—John L. Schwartz, M.D., Editor in Chief, *Psychiatric Times.*

Penguin 464 pp 0-14-025489-7 $16.00

Max Weber

### THE PROTESTANT ETHIC AND THE SPIRIT OF CAPITALISM

**and Other Writings**

*Translated and Edited with an Introduction and Notes by Peter Baehr and Gordon C. Wells.* A new translation of the original 1905 German text and the first to include Weber's unexpurgated responses to his critics.

Penguin Classics/20C

384 pp 0-14-043921-8 $16.00

Sun-tzu

### THE ART OF WAR

**The Essential Translation of the Classic Book of Life**

*Newly Translated with an Introduction and Commentary by John Minford.* This crisp, accessible new translation brings the seminal work to life for modern readers. Includes two chronologies; suggestions for further reading; and notes on the text, commentary, and pronunciation.

Viking 288 pp 0-670-03156-9 $24.95

*Available November 2002*

**PENGUIN PUTNAM INC.**

THE JOURNAL OF

PHILOSOPHY

SUBSCRIPTIONS

**2003**

Students $20.00
Individuals $35.00
Institutions $75.00
Postage outside the U.S. $15.00

**SPECIAL OFFER**

**one complimentary copy** of any of the following special issues with each new individual subscription

| | | |
|---|---|---|
| 2000 | *Causation* | David Lewis, Ned Hall, John Collins, L.A. Paul, Jonathan Schaffer |
| 1998 | *Having the World in View: Sellars, Kant, and Intentionality* | John McDowell |
| 1995 | *Reconciliation through the Public Use of Reason* | Jürgen Habermas |
| | *Reply to Habermas* | John Rawls |
| 1994 | *Sense, Nonsense, and the Senses: An Inquiry into the Powers of the Human Mind* | Hilary Putnam |
| 1990 | *The Structure and Content of Truth* | Donald Davidson |
| 1985 | *Well-Being, Agency, and Freedom* | Amartya Sen |
| 1980 | *Kantian Constructivism in Moral Theory* | John Rawls |
| 1977 | *Teleology Revisited* | Ernest Nagel |
| 1968 | *Ontological Relativity* | W.V. Quine |

All issues above are also available for $8.00 plus $3.00 for shipping/handling ($5.00 for international and Canadian orders). Postage for each added issue, $1.00. Send check or money order to:

*Journal of Philosophy*
709 Philosophy Hall
1150 Amsterdam Avenue, MC 4972
Columbia University
New York, NY 10027
(212) 666-4419
www.journalofphilosophy.org

# THE JOURNAL OF PHILOSOPHY

31/12/02

VOLUME XCIX, NUMBER 11
NOVEMBER 2002

page
553 *Taking Luck Seriously* Michael J. Zimmerman
577 *Value-pluralist Egalitarianism* Alan Carter

BOOK REVIEWS
600 Brian Barry: *Culture and Equality* Samuel Freeman

*Published by* **The Journal of Philosophy, Inc.**

# THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J E WOODBRIDGE AND WENDELL T BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

*Editors*: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.
You may also visit our website at: www.journalofphilosophy.org

# THE JOURNAL OF PHILOSOPHY

2003

SUBSCRIPTIONS (12 issues)

| | |
|---|---|
| Individuals | $35.00 |
| Libraries and Institutions | $75.00 |
| Students, retired/unemployed philosophers | $20.00 |
| Postage outside the U.S. | $15.00 |

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Ephrata, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by writing to us.

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.
POSTMASTER: Send address changes to the *Journal of Philosophy* at MC 4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2002 by the Journal of Philosophy, Inc. ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY
VOLUME XCIX, NO. 11, NOVEMBER 2002

31 12/02

# TAKING LUCK SERIOUSLY[*]

Suppose someone were to say to you, "Look, I grant that moral responsibility requires freedom and that freedom requires alternate possibilities. Nonetheless, it's perfectly possible for someone to be morally responsible even in the absence of alternate possibilities." You would be mystified. You would, in G. E. Moore's[1] gentle phrase, "be entitled to laugh at him and to distrust his future statements" about moral responsibility (*ibid.*, p. 13). So, too, if he were to say, "I grant that moral responsibility requires freedom and that freedom is incompatible with causal determinism. Still, it's perfectly possible for someone to be morally responsible even if determinism is true."

Well, I do not want you to laugh at me or to distrust my future statements about moral responsibility, and so I shall not ask you to accept either of the foregoing positions. But I shall be urging you to accept something very close to them. As a corollary, I shall also be

---

[*] Earlier versions of this article were presented at the fourth annual Inland Northwest Philosophy Conference, at a meeting of the Research Triangle Ethics Circle, and at the Universities of Gothenburg, Lund, Massachusetts, Oxford, Stockholm, Umeå, and Uppsala. I am grateful for the comments of many people, including Robert Allen, Gustaf Arrhenius, Gunnar Bjornsson, Ben Bradley, John Broome, Allen Buchanan, Krister Bykvist, Joe Campbell, Erik Carlson, Randy Clarke, Sven Danielsson, John Davis, John Eriksson, Richard Feldman, Bob Frazier, Bruce Galbreath, Don Garrett, Bernard Gert, Carl Ginet, Ish Haji, Chris Heathwood, Tom Hill, Magnus Jiborn, Jens Johansson, Bob Kane, Sten Lindström, Ruth Lucier, Hans Mathlein, Terry McConnell, Sean McKeever, Christian Munthe, Stuart Opotowsky, Jan Österberg, Ingmar Persson, Tomasz Pol, Wlodek Rabinowicz, Joseph Raz, Toni Rønnow-Rasmussen, Gary Rosenkrantz, Lars Samuelsson, Geoff Sayre-McCord, Ted Sider, Howard Sobel, Jim Sterba, Bertil Strömberg, Frans Svensson, Torbjörn Tännsjö, Folke Tersman, Peter van Inwagen, Bernard Williams, and David Wong.

[1] *Principia Ethica* (New York: Cambridge, 1903)

0022-362X/02/9911/553–76

urging you to rethink your stand on several of our current practices, including especially the practice of punishment.

Let me begin by distinguishing judgments about moral responsibility from two other types of judgments with which they are apt to be confused.

## I

First, judgments about moral responsibility are distinct from judgments about moral right, wrong, and obligation. Judgments of the latter sort are frequently called *deontic*. There is no standard term to refer to judgments about moral responsibility, and so let me coin one: *hypological*.[2] Hypological judgments have (primarily) to do with the moral praiseworthiness (or laudability) and blameworthiness (or culpability) of persons; they constitute one type of agent evaluation. Deontic judgments are quite different. It is common to say that deontic judgments constitute a type of act evaluation, but I am not sure that this is correct. We do admittedly say things like "What Joe did was right (wrong, obligatory)," and this seems to constitute an evaluation of Joe's act. But we also say things like "Joe was right (wrong, obligated) to do what he did," and this seems to constitute an evaluation of Joe. But, even if in the end we should declare deontic judgments a species of agent evaluation, they are quite different from hypological judgments. It is a commonplace that one can do right (or wrong) without being praiseworthy (or blameworthy). It is less commonly recognized, but nonetheless true (I believe[3]), that one can be praiseworthy (or blameworthy) without doing right (or wrong).

Second, judgments about moral responsibility are distinct from judgments about moral virtue and vice. Judgments of the latter sort—often called *aretaic*—constitute a type of character evaluation. The relationship between persons and their characters is admittedly complex, and I do not propose to investigate the matter here. But, despite the fact that we often say things like "Joe is quite a character," I think it is pretty clear that persons should not be said to be characters so much as to *have* characters. Again, though, even if in the end we should declare aretaic judgments a species of agent evaluation, they are quite different from hypological judgments. It is a commonplace that one can have a certain character trait that has no bearing on one's moral responsibility. It is less commonly recognized, but nonetheless true (I believe[4]), that one can be morally responsible for something in a way that has no bearing on one's character.

---

[2] I draw this term from the Greek ὑπόλογος, meaning "held accountable or liable."

[3] I argue for this in *An Essay on Moral Responsibility* (Totowa, NJ: Rowman and Littlefield, 1988), chapter 3, and in "A Plea for Accuses," *American Philosophical Quarterly*, xxxiv, 2 (April 1997): 229-43.

Even among those who agree that hypological judgments are distinct from both deontic and aretaic judgments, there is, however, disagreement about just what such judgments are judgments *of*. There are two closely related views on this which I shall mention here. The first is that to be morally responsible just is to be the appropriate object of one or more of what P. F. Strawson[5] calls the "reactive attitudes," such as resentment, gratitude, forgiveness, and the like. This view has recently garnered considerable support.[6] The second view is that to be morally responsible is to be such that there is an "entry" in one's "moral ledger" in light of some fact about oneself; one's "moral record as a person" is affected by this fact.[7] (In putting matters this way, I do not mean to presuppose either that it is or that it is not possible to arrive at an overall assessment of one's moral worth by somehow aggregating the individual entries in one's ledger.) This is to put the second view only roughly, for a person may have a number of moral ledgers or records; but it can be made more precise by allying it to some degree with the first view. The moral record at issue is precisely that which renders the person the appropriate object of reactive attitudes (and, moreover, liable to more robust reactive measures, such as reward and punishment, that incorporate but extend beyond such attitudes). The difference between the first and second views is that, whereas the former identifies responsibility with susceptibility to certain reactive attitudes, the latter identifies responsibility with that in virtue of which one is susceptible to such attitudes. On the first view, the claim that responsibility is the proper occasion for certain reactive attitudes is analytic; on the second view, this is a substantive claim whose truth can be sensibly disputed.[8]

---

[4] I argue for this in *An Essay on Moral Responsibility*, chapter 4, section 4.8.

[5] "Freedom and Resentment," *Proceedings of the British Academy*, XLVIII (1962): 187-211.

[6] Cf. R. Jay Wallace, *Responsibility and the Moral Sentiments* (Cambridge: Harvard, 1994); David Copp, "Defending the Principle of Alternate Possibilities: Blameworthiness and Moral Responsibility," *Noûs*, XXXI, 4 (December 1997): 441-56; and John Martin Fischer and Mark Ravizza, *Responsibility and Control: A Theory of Moral Responsibility* (New York: Cambridge, 1998).

[7] Cf. Joel Feinberg, *Doing and Deserving* (Princeton: University Press, 1970); Jonathan Glover, *Responsibility* (New York: Humanities, 1970); Herbert Morris, *On Guilt and Innocence* (Berkeley: California UP, 1976); and my *An Essay on Moral Responsibility*.

[8] It can be disputed from both ends. It might be claimed that it is possible to be responsible without being the proper target of any reactive attitudes, and it might be claimed that it is possible to be the proper target of some reactive attitude

Although nothing will turn on this here, it is the second view to which I subscribe and in terms of which I shall couch my thesis. When I say that a person is praiseworthy, I shall mean that her moral record is favorably affected by some fact about herself; when I say that a person is blameworthy, I shall mean that her moral record is adversely affected by some such fact. To praise or blame someone, in this sense, is simply to make a judgment about her moral record, a judgment which may form the basis of, but which is not itself, a "reaction" either in attitude or in some more robust form of behavior toward that person.

It is standardly acknowledged that there are two key components of moral responsibility, one epistemic and the other metaphysical. Here, I shall concentrate on the latter, which has to do with the freedom or control that the agent enjoys.[9] Almost all writers on the subject have assumed that moral responsibility presupposes some form of freedom or control. There have of course been exceptions. Robert Merrihew Adams,[10] for example, has argued that we are responsible for our sins, whether or not they are voluntary. The graduate of the Hitler *Jugend*, he says, is to be blamed for his beliefs and actions, regardless of whether they are in his control (*ibid.*, p. 19). Similarly, Eugene Schlossberger[11] has claimed that moral responsibility is simply a matter of moral evaluability, and there is no requirement that the agent enjoy any measure of control over that for which he is evaluable. One may be properly morally evaluated, Schlossberger says, in light of not just one's actions but also one's beliefs, emotions, and so on, irrespective of whether one is in control of them (*ibid.*, pp. 6ff., 37ff., 101ff.). Thomas Scanlon[12] contends that one is morally respon-

---

without being responsible. Whether such claims succeed depends on what is meant by 'proper' We should distinguish between someone's being morally deserving of a certain reaction and its being overall morally justifiable to react to that person in that way. When I talk of "susceptibility" to reactive attitudes, it is the deservingness of such reactions that I mean. I concede that it is possible for such a reaction to be deserved but unjustified (because, say, no one has the proper authority to react in the prescribed manner—see Feinberg, p. 128), and that it is possible for such a reaction to be justified but undeserved (as proponents of strict liability in effect contend). What I deny is that it is possible that someone be morally responsible but not deserving of one of the reactive attitudes, or vice versa. But I shall not seek to defend this claim here.

[9] Some writers contend that authenticity is a necessary condition of moral responsibility and that it is distinct from both the epistemic component and the control component. See, for example, Ishtiyaque Haji, *Moral Appraisability: Puzzles, Proposals, and Perplexities* (New York: Oxford, 1998), chapters 6-7.

[10] "Involuntary Sins," *Philosophical Review*, xciv, 1 (January 1985): 3-31.

[11] *Moral Responsibility and Persons* (Philadelphia: Temple, 1992).

[12] *What We Owe to Each Other* (Cambridge: Harvard, 1998)

sible for having certain attributes, even if one was not in control of coming to possess them, so long as it is appropriate to ask that one defend or disown them (*ibid.*, pp. 274ff.). And others[13] have made similar claims.

If the view of Adams and others were correct, then the perennial philosophical preoccupation with the relation between freedom and causal determinism would be badly misguided, insofar as it is driven (as it very often is) by a concern with coming to terms with moral responsibility. But I do not think that it is correct. Adams and others are, I believe, quite right to say that not all moral evaluations of or concerning an agent presuppose that the agent is in control of that in light of which the evaluations are made. It is surely correct to say that there can be, as Michael Slote[14] puts it, ethics without free will. Much of virtue ethics, I would say, has nothing whatever to do with freedom or control. But that, of course, does not mean that the sort of moral evaluation that has to do with moral responsibility in particular has nothing to do with freedom or control. For example, while we may justifiably condemn evil wherever we find it, the fact is that it comes in various forms. In a recent, interesting discussion of moral evil, Daniel M. Haybron[15] contrasts the sort of evil embodied by Claggart, the master-at-arms in Herman Melville's *Billy Budd*, with the sort of evil embodied by Dorian Gray in Oscar Wilde's story. He puts the matter well:

> The purely evil individual [such as Claggart] is unquestionably vile, but he lacks an important fault: he does not give himself freely to evil, but is delivered to it. Claggart could not help but be a cruel man—that's just the way he *is*. Dorian Gray's cruelty, on the other hand, is entirely of his own making (*ibid.*, p. 143).

On the assumption that Haybron's observation is accurate, Claggart, though clearly morally evaluable in light of his particular brand of cruelty, is not, I would say, morally responsible for it, whereas this cannot be said of Dorian Gray regarding the cruelty that characterizes him.

---

[13] For example, Gary Watson, "Two Faces of Responsibility," *Philosophical Topics*, xxiv, 2 (Fall 1996). 227-48, pp. 233-34. Watson is here discussing that type of moral responsibility that he calls "attributability," which is at least very close to the conception of responsibility that is at issue here. This is also the term that Scanlon uses to refer to the type of responsibility under discussion.

[14] "Ethics without Free Will," *Social Theory and Practice*, xvi, 3 (Fall 1990): 369-83. Cf my *An Essay on Moral Responsibility*, chapter 4, section 4.9; and Michael Otsuka, "Incompatibilism and the Avoidability of Blame," *Ethics*, cviii, 4 (July 1998): 685-701, pp. 694-95.

[15] "Evil Characters," *American Philosophical Quarterly*, xxxvi, 2 (April 1999): 137-48.

There are complications, of course. We should distinguish between *having* dispositions and *acting* on them, and so too between having control over and being morally responsible for one's dispositions and having control over and being morally responsible for acting on them. Still, at this point, let me just say that I side with the majority who declare freedom or control essential to moral responsibility.[16] Thus the relation between freedom and causal determinism is for me a live issue. What I want to argue here is that a great deal of recent discussion, enlightening though much of it has been, has missed what is crucial about the connection between freedom and responsibility. Once what is crucial has been correctly identified, something strange happens: the connection itself becomes quite tenuous. This, in turn, threatens to undermine many of our common practices—in particular, the practice of punishment.

## II

The great bulk of recent discussion of the connection between freedom and moral responsibility has focused on the issue of whether the control required by moral responsibility itself requires that the agent have the option to choose or act in a manner distinct from that in which he does choose and act—the issue, that is, of whether moral responsibility is possible in the absence of alternate possibilities. Harry G. Frankfurt[17] has famously argued that moral responsibility does *not* require alternate possibilities. If correct, this is obviously important, inasmuch as the traditional incompatibilist position is one according to which moral responsibility is incompatible with causal determinism precisely because such responsibility requires alternate possibilities and alternate possibilities are incompatible with causal determinism. There are several[18] who have rejected Frankfurt's argument, and their voice has grown increasingly strong in recent years.

---

[16] This is not to say that holding someone strictly liable for an offense cannot be morally justified. See footnote 8 above.

[17] "Alternate Possibilities and Moral Responsibility," this JOURNAL, LXVI, 23 (December 4, 1969): 829-39.

[18] For example: David Blumenfeld, "The Principle of Alternate Possibilities," this JOURNAL, LXVIII, 11 (June 1971): 339-45; Peter van Inwagen, *An Essay on Free Will* (New York: Oxford, 1983), pp. 161ff.; Margery Bedford Naylor, "Frankfurt on the Principle of Alternate Possibilities," *Philosophical Studies*, XLVI, 2 (September 1984): 249-58; David Widerker, "Libertarianism and Frankfurt's Attack on the Principle of Alternate Possibilities," *Philosophical Review*, CIV, 2 (April 1995): 247-61; Carl Ginet, "In Defense of the Principle of Alternative Possibilities: Why I Don't Find Frankfurt's Arguments Convincing," *Philosophical Perspectives*, X (1996): 403-17; Robert Kane, *The Significance of Free Will* (New York: Oxford, 1996); Copp; Keith Wyma, "Moral Responsibility and Leeway for Action," *American Philosophical Quarterly*, XXXIV, 1 (January 1997): 57-70; Otsuka; and Scott A. Davison, "Moral Luck and the Flicker of Freedom," *American Philosophical Quarterly*, XXXVI, 3 (July 1999): 241-51.

There are also several[19] who have accepted it. The discussion has been complex and subtle, and I believe much progress has been made. Nonetheless, all parties to the debate have tended to overlook what is crucial to the relation between freedom and moral responsibility.

Why is freedom commonly regarded as important to moral responsibility? The usual answer is simply that we cannot be morally responsible for what is not in our control.[20] I believe that this answer is accurate but incomplete. There is a more general point to be made, and that is that the degree to which we are morally responsible cannot be affected by what is not in our control. Put more pithily: luck is irrelevant to moral responsibility.[21] This is the crucial point at issue. Pursuit of it will show that the question whether freedom requires alternate possibilities, and even the question whether freedom is compatible with causal determinism, paradoxically become considerably less significant than they are frequently taken to be.

The relevance of luck to moral responsibility has been widely debated ever since the publication of the influential pair of papers on moral luck written by Bernard Williams and Thomas Nagel.[22] But seldom have the implications of the denial of the relevance of luck to moral responsibility been pursued to their logical conclusion. It is this that I shall undertake here. We can distinguish between two broad types of luck, which I shall call *resultant* and *situational*.[23] The former consists in luck with respect to the results of one's choices and actions; the latter consists in luck with respect to the situations in which one finds oneself. It is the former that is most often discussed, but it is the latter whose implications run deepest and are, as we shall see, especially subversive of our everyday judgments about moral responsibility.

---

[19] For example: Fischer, *The Metaphysics of Free Will. An Essay on Control* (Cambridge: Blackwell, 1994); Haji; Alfred Mele and David Robb, "Rescuing Frankfurt-Style Cases," *Philosophical Review*, CVII, 1 (January 1998): 97-112; and David Hunt, "Moral Responsibility and Avoidable Action," *Philosophical Studies*, XCVII, 2 (January 2000): 195-227.

[20] As is commonly recognized, the control in question may be merely remote rather than immediate. The standard case of the drunk driver, who was in control of his drinking but is not in control of his driving, establishes this point.

[21] Clearly, I am here using 'luck' as follows: something that occurs as a matter of luck is something that occurs beyond one's control. I am not using it, as some do (for example, Nicholas Rescher, "Luck," *Proceedings and Addresses of the American Philosophical Association*, LXIV, 3 (November 1990): 5-19), as follows: something that occurs as a matter of luck is something that occurs by chance, that is, something that is such that there is or was some probability of its not occurring.

[22] Williams, "Moral Luck," *Proceedings of the Aristotelian Society*, Supplementary Volume L (1976): 115-35; Nagel, "Moral Luck," *Proceedings of the Aristotelian Society*, Supplementary Volume L (1976): 137-51.

[23] These terms were introduced in my "Luck and Moral Responsibility," *Ethics*, XCVII, 2 (January 1987): 374-86.

## III

We are all familiar with resultant luck. Examples of it abound in the literature. Here is one. "How is it possible," Nagel writes, "to be more or less culpable depending on whether...a bird [gets] into the path of one's bullet" (*op. cit.*, p. 143)? My answer is that this is not possible. Let us compare cases. Suppose that George shot at Henry and killed him. Suppose that Georg shot at Henrik in circumstances which were, to the extent possible, exactly like those of George (by which I mean to include what went on "inside" the protagonists' heads as well as what happened in the "outside" world), except for the fact that Georg's bullet was intercepted by a passing bird (a rather large and solid bird) and Henrik escaped injury. Inasmuch as the bird's flight was not in Georg's control, the thesis that luck is irrelevant to moral responsibility implies that George and Georg are equally morally responsible. This, I believe, is absolutely correct.

You may have some doubts. "If George and Georg are equally responsible," you may say, "then, since Georg is not responsible for killing Henrik—the bird luckily got in the way (or unluckily, depending on whose perspective is at issue: Henrik's, Georg's, or the bird's)—it follows that George is not responsible for killing Henry. But that's absurd. If that were the case, no one would ever be responsible for killing anyone, since success in one's endeavors always requires the cooperation of factors that are beyond one's control."

This objection is based on a misunderstanding. I do not wish to deny that George is responsible for killing Henry (or for Henry's death—the distinction between actions and their "results"[24] seems to me irrelevant here); whether he is so will depend on details of the case (both epistemic and metaphysical) which I have not supplied. And, of course, I concede that Georg is not responsible for killing Henrik (or for Henrik's death), since Henrik did not die. Thus I am quite willing to grant that George may well be *responsible for more things* than Georg. What I deny is that George is any *more responsible* than Georg. We must distinguish the *degree* of someone's responsibility from its *scope*. (The term 'extent' strikes me as ambiguous between the two.) My claim is that George and Georg bear responsibility to the same degree, despite the fact that George's responsibility has greater scope. Let us suppose that George committed murder when he killed Henry and is indeed to blame for doing so. (This requires that George satisfy both some epistemic condition and some condition concerning freedom or control. There is no need for me to specify these conditions further; I invite you to fill in the details as you deem fit.) On the view of responsibility adumbrated earlier, this means that George's moral record as a person is adversely affected in some way in virtue of the fact that he killed Henry as he did. My claim is that, although Henrik survived Georg's attempt to kill him, Georg's moral record as a person is adversely affected in precisely the same way.

---

[24] To use a term introduced by Georg Henrik von Wright—who is neither the Georg nor the Henrik of my example—in *Explanation and Understanding* (Ithaca: Cornell, 1971), p. 66.

"But what," you may ask, "is Georg supposed to be responsible for? In virtue of what is his moral record supposed to be adversely affected?" The answer is simple: he is responsible for his attempt on Henrik's life, just as George is responsible for his attempt on Henry's life. The fact that Georg's attempt was unsuccessful, whereas George's was successful, is irrelevant to the question of how blameworthy they are.

"But that is to trivialize Henry's death," you may protest. No, it is not. Clearly, something terrible happened when George killed Henry, something that has no counterpart in the case of Georg and Henrik. It may even be agreed that George did something morally wrong that Georg did not. But that is a *deontic* judgment.[25] My claim is simply that, when it comes to judgments about *responsibility*, more particularly to judgments about *degrees* of responsibility, George and Georg are on a par. It is especially important to note that this claim affords Georg no excuse whatsoever. I have said that George is no more to blame than Georg, and that may seem to suggest that George is not particularly blameworthy. But, of course, there is no such implication. I could equally well have said that Georg is no less to blame than George; the passing bird in no way mitigates Georg's blameworthiness.

You may still be uneasy. "Wouldn't it be appropriate," you may ask, "to react more harshly toward George than toward Georg—for instance, to punish him more severely? Doesn't this show that George is more blameworthy than Georg after all?" This is a difficult matter. I have agreed that responsibility is directly correlated, even if it is not identical, with susceptibility to reactive attitudes (and to more robust reactive measures). Given this, it might seem that I am committed to denying that it is appropriate to react more harshly toward George than toward Georg. And in one sense I am. I am committed to denying that George *deserves* (in virtue of his moral record) a harsher reaction than Georg. But that leaves open the possibility that it would be morally justified to react more harshly toward George than toward Georg for reasons *other* than those having to do with desert (or, more precisely, desert rooted in George's moral record). Perhaps there are good moral reasons to punish murder, such as that committed by George, more severely than a failed attempt at murder, such as that committed by Georg—I am not sure about this; but, if there are, they are grounded in something other than relative degree of responsibility.[26]

---

[25] This judgment is based on the actual consequence (or result) of George's action. Someone who takes right and wrong to be a function of foreseeable consequences might reject the judgment.

I believe that anyone who takes seriously the view that we cannot be morally responsible for that which is not in our control must acknowledge that George and Georg are equally responsible and thus accept the more general claim that luck is irrelevant to moral responsibility. I recognize that there is a clear sense in which George was in control of Henry's death; he was (we may assume) in control of whether he shot at Henry in the way that he did, and, under the circumstances, shooting at Henry in that way was all that was needed to kill him. He shot at Henry, and Henry consequently died; had he not shot at Henry, Henry would not have died. Nonetheless, it is also true that George controlled Henry's death only to the extent that he controlled his shooting at Henry; the other factors that conspired to produce Henry's death were not in his control at all. In this respect, George was no more in control of what happened to Henry than Georg was in control of what happened to Henrik.[27] Just as with responsibility, so too with control: we must distinguish degree from scope. George was *in control of more things* than Georg (his control had greater scope), but he was no *more in control* of what happened than Georg was (he was in control to the same degree). Insofar as degree of responsibility tracks degree of control, George and Georg must be declared equally morally responsible.

Although what I have said about equality of desert already casts some doubt on our current practice of punishment, there is nothing in what I have said so far that impugns the significance of the questions whether freedom requires alternate possibilities and whether freedom is compatible with causal determinism. If George and Georg are morally responsible for their respective attempts at murder, this is (in part) because they freely committed these attempts. Whether such freedom requires that they could have acted differently seems an important question. Whether such freedom requires causal indeterminism likewise seems an important question. The thesis that resultant luck is irrelevant to moral responsibility does nothing to reduce the urgency of these questions.

---

[26] Cf. footnote 8 above.

[27] We must therefore accept that, in the pertinent sense of 'control', an agent may have control over something even though that thing would not occur or obtain but for something else over which the agent has no control. Just as Georg's failure to kill Henrik is a matter of luck, so too is George's success in killing Henry. To a certain, indeed a very great, extent, luck is ineliminable from our lives; but that does not mean that we can exercise no control whatsoever.

## IV

But the thesis that situational luck is irrelevant to moral responsibility is a different matter. Once again, this is a type of luck with which we are all familiar. But few[28] appear to take it seriously, perhaps because doing so is disturbingly humbling. "There but for the grace of God go I," we may mutter on occasion, but then we quickly turn away from such a discomfiting thought.

There are two basic varieties of luck regarding the situations in which one finds oneself: the first having to do with one's *circumstances*, the second with one's *constitution*. I shall discuss each in turn.

Return to George and Henry and their counterparts, Georg and Henrik. Suppose, as before, that George shot at Henry and killed him. Suppose also, as before, that Georg did not kill Henrik; suppose now, however, that this was not because he took a shot that was intercepted by some unfortunate bird, but rather because he took no shot at all. And suppose that this was because of something quite fortuitous: Georg sneezed just as he was about to shoot, for example; or a truck pulled up in front of Henrik, blocking Georg's line of fire; or Henrik turned suddenly into a doorway, just as Georg was about to squeeze the trigger. Whereas in the case involving the bird, luck intervened after the shot took place, thereby preventing Henrik's death, in this sort of case the intervention occurs earlier, before Georg has a chance to act at all. But the cases are united in that, in all of them, Georg would have freely killed Henrik but for some feature of the case over which he had no control. This being so, it seems that we must conclude here, as before, that Georg is as culpable as George. The circumstances that conspired to save Henrik afford Georg no excuse.

Again, you may have some doubts. "Excuse for what?" you may ask. "What is Georg supposed to be responsible for this time? In this sort of case, there isn't even an attempt on Henrik's life that you can pin on Georg." One answer to this question is simply to say that what Georg is responsible for is his being such that he would have freely killed Henrik, given the opportunity. But actually I find this quite dubious. I have said that I subscribe to the view that we cannot be responsible for what is not in our control, and I doubt it should be said that Georg was in control of his being such that he would have freely killed Henrik.[29] In my view (which I shall not try to defend here), an agent exercises control directly over his choices (that is, his choosings) and indirectly over the consequences of his choices. In the sort of case under discussion, Georg's being such that he would have freely killed Henrik is clearly not itself a choice of his; nor is it the consequence of a choice of his. Rather, he would have freely killed Henrik because he *would* have freely chosen to shoot him, had he had the cooperation of certain features of the case; and this is a choice which did *not* occur, precisely because the requisite cooperation was not forthcoming.

---

[28] Among philosophers, Feinberg (*op. cit.*), and Nagel (*op. cit.*), are prominent exceptions.

I think, therefore, that we should reject the claim that Georg is responsible for being such that he would have freely killed Henrik. And I think that there is nothing else for which Georg might be said to be responsible.[30] If so, this sort of case serves to emphasize, in even more dramatic fashion than before, the distinction between degree and scope of responsibility. The degree of Georg's responsibility remains the same as George's, but the scope of Georg's responsibility has dwindled to nothing. Georg is responsible; he is just not responsible *for* anything. He is, as I shall put it, "responsible *tout court*." Lest this appear unduly paradoxical, let me hasten to add that it is nonetheless the case that Georg is responsible *in virtue of* something, and this something just is his being such that he would have freely killed Henrik, had he had the cooperation of certain features of the case.[31] All responsibility, including responsibility *tout court*, is fundamentally relational. It is precisely because George is responsible in virtue of the very same sort of fact (the fact that he would have freely killed someone, had he had the cooperation—as he did—of certain features of the case) in virtue of which Georg is responsible, that George and Georg are responsible to the same degree.

---

[29] This is not to say that it cannot happen that one is in control of such a fact. On the contrary, it can. See footnote 33 below.

[30] You might be tempted to say that Georg is responsible for the *intention* to kill someone, and that it is this that renders him relevantly like George. I concede that this is possible, but I submit that it may well not be the case. It frequently happens that we are not in control of the intentions that we form. In such cases, indeed, it is rather misleading to talk of "our forming the intentions"; better would be to talk of "the intentions forming in us." Moreover, it could of course happen that luck intervenes *prior* to the formation of an intention. (In Feinberg's original, memorable examples (*op. cit.*, p. 35), sneezes and loud noises intrude at just such a point.)

You might also be tempted to say that Georg is responsible for all the preparatory actions he performed prior to his aborted attempt on Henrik's life. Again, I concede that this is possible, even likely. But note two points. First, it is not necessary. It is conceivable (barely) that Georg undertook no preparations; it is conceivable (easily) that he undertook preparations for which he is, for one reason or another, not responsible. Second, and more importantly, even if Georg undertook preparations for which he is responsible, this does not exhaust his responsibility. Remember that the cases of George and Georg are being assumed to be, as far as possible, exactly alike, but for the fact that George took a shot and Georg did not. Thus their preparations were the same. George's responsibility is not limited to his preparations; hence Georg's is not limited to his, either.

Let me turn now from *circumstantial* to *constitutive* luck,[32] from luck having to do with one's *external situation* to luck having to do with one's *internal situation*. Suppose, for instance, that the reason why Georg did not kill Henrik was that he was too timid, or that he had a thick skin and Henrik's insults did not upset him in the way that Henry's insults upset George, or that he was deaf and simply did not hear the insults that Henrik hurled his way. If it is nonetheless true that Georg *would* have freely shot and killed Henrik but for some such feature of the case over which he had no control, then, I contend, he is just as responsible, in virtue of this fact, as George is.[33]

---

[31] Contrast Copp (pp. 449-50), where being responsible for something and being responsible on the basis of something are identified.

[32] I borrow these terms from Nagel.

[33] It is sometimes maintained that taking luck seriously, in the manner that I have advocated, culminates in the view that no one is ever responsible. (See, for example, Michael Moore, *Placing Blame* (New York: Oxford, 1997), pp. 242–43.) As I have just shown, this is not so.

The general principle upon which I have relied in my argument may be put as follows: if (a) someone's being $F$ (where '$F$' designates some complex property comprising both epistemic and metaphysical components) is sufficient for that person's being morally responsible to some degree $x$, then, if (b) it is true of $S$ at some time that he or she would be $F$ if $p$ were true, and (c) $p$'s being true is not in $S$'s control at that time, then (d) $S$ is morally responsible to degree $x$. You may think this principle too liberal. Suppose that Dave would have freely performed some heroic act if he had not been drunk at the time and that, by that time, his being drunk was not in his control. Should we really judge him to be as laudable as he would have been had he been sober and acted heroically? I am inclined to think that we should. Suppose that his being drunk was not his fault (the whisky was forced down his throat, as with Cary Grant in *North by Northwest*); then, just as Georg does not escape culpability due to luck, so too Dave should not be thought to "escape" laudability due to luck. Suppose that his being drunk was Dave's fault; then that is certainly part of his moral record, but I do not see why this should preclude his counterfactual heroism's also being part of his moral record. Still, I need not insist on this. If you like, you may qualify clause (d) of the principle I have just given as follows: unless $S$ is morally responsible for the fact that (c) $p$'s being true is not in $S$'s control at that time. This would cover all the cases I have presented, since they all implicitly presuppose that Georg is not morally responsible for the fact that he was not in control of the relevant feature that prevented Henrik's death.

## V

If what I have just said is true, then I believe the question whether freedom requires alternate possibilities loses a good deal of its significance. This is because even those who deny that freedom requires alternate possibilities will, of course, agree that freedom does not require the *absence* of alternate possibilities. So let us suppose that, when George freely killed Henry, he had the option not to do so. Given that he satisfied whatever epistemic requirement must be met in order to be responsible, everyone will agree that George is responsible for killing Henry.[34] What should we say if we were now to suppose that George lacked the option not to kill Henry, due to some Frankfurtian constraint? Some would say that George freely killed Henry anyway; others would say that, if the constraint really did deprive George of any (relevant) alternate possibility, then he did not freely kill Henry after all. Given the assumption that moral responsibility requires freedom, many have understandably thought that it is therefore very important to determine whether the agent does act freely in such a case. But, as I see it, this concern is considerably diluted, if it is agreed that George would have freely killed Henry in the absence of the Frankfurtian constraint. For we should then conclude that, regardless of whether he did freely kill Henry, George is in fact just as responsible as he would have been had he freely killed Henry.

Indeed, in light of this, even the supposedly fundamental question whether freedom is compatible with causal determinism loses much of its force. (It may retain more of its force than the question regarding alternate possibilities, however, insofar as some of those who deny that freedom requires indeterminism do claim that it requires determinism.[35]) Consider a case of luck that has to do with

---

You may think that this issue could be circumvented by reformulating clause (c) of the original principle as follows: *p*'s being true is not in *S*'s control at *any* time. Although I believe this revised principle to be true, it is too restrictive for my purposes. For instance, it would not cover the case in which Georg sneezes. Although Georg was not in control of his sneezing at the time he sneezed, there was a time at which he was in control of it. He could have shot and killed himself the day before; that would certainly have inhibited the sneeze.

[34] This is somewhat overstated. See footnote 9 above

[35] Cf. R. E. Hobart, "Free Will as Involving Determination and Inconceivable without It," *Mind*, XLIII, 169 (January 1934): 1-27. There are, of course, still others who argue that freedom is impossible, on the grounds that it is compatible with neither determinism nor indeterminism. Cf. Galen Strawson, *Freedom and Belief* (New York: Oxford, 1986); Richard Double, *The Non-Reality of Free Will* (New York: Oxford, 1991).

what may be called *antecedent circumstances*.[36] Suppose that Georg had been subjected to some *Clockwork Orange*-type conditioning process that rendered him incapable of killing. (This would seem to count as a case of constitutive as well as circumstantial luck.) If it is nonetheless true that he would have freely killed Henrik, but for this conditioning process over which he had no control, then he is just as responsible as George. Suppose, now, that Georg was not *conditioned* not to kill Henrik but simply that he was *deterministically caused* not to (the chain of causation extending back before Georg's birth, if you wish). Incompatibilists would claim that his not killing Henrik was therefore unfree, whereas compatibilists would deny this. But, as I see it, the significance of this dispute is considerably reduced by the observation that, even if the incompatibilist is right, Georg is still as responsible as George, if he would have freely killed Henrik, had his causal history cooperated.[37] My conclusion is that it is perfectly possible for someone to be morally responsible, even if causal determinism is true and even if freedom is incompatible with such determinism. To put it bluntly, it does not matter whether Georg *could* have killed Henrik. What matters is whether he *would* have freely killed him, *if* he had the cooperation of certain features of the case. And I say this even though I take freedom to be crucial to judgments of moral responsibility.

To get clear on just what my proposal is, let us return for a moment to the two positions mentioned at the outset. They can be laid out formally as follows:

(A) (1) Moral responsibility requires freedom.
(2) Freedom requires alternate possibilities.
(3) Moral responsibility does not require alternate possibilities.
(B) (1) Moral responsibility requires freedom.
(2) Freedom requires causal indeterminism.
(3) Moral responsibility does not require causal indeterminism.

In each case, we have an inconsistent triad. To resolve the inconsistency, what I have proposed is this. First, I have conceded that one cannot be morally responsible *for* something unless that thing is or was in one's control, and so in this sense clause (1) of (A) and (B) must be granted.[38] If clause (2) should also be granted (a matter about which I have been noncommittal; note that it is possible that clause (2) should be granted in one case but not the other), then of course clause (3) must be rejected. But I have argued, further, that one can be morally responsible *tout court*, that is, responsible without being responsible *for* something, and that, on this understanding, clause (1) of (A) and (B) should be rejected. Thus, even if clause (2) should in each case be granted, clause (3) may (and, I believe, should) also be accepted. That is precisely what reduces the significance of the question whether clause (2) is true.

---

[36] Again, the term is Nagel's (*op. cit.*).

[37] I concede that, if incompatibilism is true, it is not at all clear just how in practice to assess the truth value of such a counterfactual. But that such a statement can in principle be true seems very plausible

[38] Again, the control may be merely remote; see footnote 20 above. It would, in fact, be consistent with the general spirit of my thesis to allow an exception to the claim that moral responsibility for something requires control over that thing.

At this point I should say something to forestall certain possible misunderstandings.

First, do not be misled by what I have just said into thinking that I am invoking two types of moral responsibility here. On the contrary, there is just one type. George and Georg are to be morally evaluated in exactly the same way, even though George is responsible for something that Georg is not. They are equally responsible; if George is deserving of a particular reaction, then Georg is deserving of the very same reaction. This indicates that whether there is something *for* which one is responsible is immaterial; all that matters, fundamentally, is whether one *is* responsible. Degree of responsibility counts for everything, scope for nothing, when it comes to such moral evaluation of agents. Thus my concession just now that clause (3) of (A) and (B) may have to be rejected, if one's concern is with the possibility of one's being responsible *for* something, is of small moment. What matters, at bottom, is that clause (3) may and, I believe, should be accepted, if one's concern is with responsibility *tout court*.[39]

---

Suppose that some event occurs over which an agent lacks control but which is such that he would have freely brought it about had he had the opportunity to do so. In such a case, I suppose we could say not only that the agent (given that he satisfies whatever epistemic condition is necessary for responsibility) is responsible in virtue of being such that he would have freely brought the event about, but also that he is responsible for this event. But although we could say this, I see no advantage to doing so and one disadvantage: it is messy.

[39] In "A Second Paradox concerning Responsibility and Luck," *Metaphilosophy*, XXVI, 1 and 2 (January and April 1995): 81-96, John Greco draws a distinction between what he calls moral responsibility and moral worth that closely resembles the distinction I have drawn between responsibility for something and responsibility *tout court* (p. 91). His discussion suggests, however, that he takes there to be two fundamentally different sorts of evaluation at issue in this context, whereas I have denied this. (Indeed, he seems to be thinking of moral worth—the sort of moral worth that concerns him—along aretaic rather than hypological lines.) Moreover, Greco nowhere indicates that he agrees with me that responsibility for something is immaterial, in the sense just explained. (A distant ancestor of Greco's paper bore the title "Taking Luck Seriously," which I have happily pilfered.)

Second, do not be misled by my claim that clause (1) of (A) and (B) is to be rejected (given that one's concern is with responsibility *tout court*) into thinking that I am reverting to the position of Adams and others mentioned above, to the effect that freedom is irrelevant to moral responsibility. On the contrary, I want to insist that it is pivotal, in that the degree to which we are morally responsible cannot be affected by what is not in our control. It is this crucial fact which renders luck, not freedom, irrelevant to responsibility. Granted, my view implies that Claggart might after all be as responsible as Dorian Gray; this would be true if Claggart would have freely given himself to evil, as Gray did, had he had the opportunity to do so. In saying this, however, I am not collapsing the distinction between aretaic judgments and hypological judgments upon which I insisted earlier. That distinction remains. I am saying that, in addition to the sort of aretaic judgment that is appropriate to Claggart (but not to Gray) in light of his particular type of moral depravity, it may be that a certain judgment about moral responsibility is equally appropriate to both.[40]

In saying this, I am invoking the view (which I attributed above to Slote, and with which Haybron clearly concurs in his discussion of Claggart) that some types of moral judgments, in particular certain judgments about moral virtue and vice, do not presuppose that we enjoy freedom of will. A sadist is evil, I would say, even if he cannot control either having or acting on his sadistic impulses. Naturally, this may be resisted. Such a person, it might be objected, is no more evil than a "vicious" dog (the quotation marks constituting an acknowledgment that a dog, of course, cannot really have a vice, let alone a moral one), precisely because he shares the inability to control his viciousness. But while we should surely agree that a vicious dog is not evil, there are possible explanations of this which do not appeal to its lack of control over its having or acting on its vicious impulses. One explanation is that a dog is not the sort of creature that typically has such control, whereas a sadist is. Perhaps this has some merit, but to my mind a better explanation is that a dog lacks the capacity to reflect on its behavior in moral (or morally relevant) terms, whereas a sadist typically does not.

Suppose I am wrong about this. Suppose that luck is as irrelevant to aretaic as to hypological judgments. Suppose, further, that nothing else serves to distinguish judgments about virtue and vice from judgments about responsibility *tout court*.[41] In this case, my argument implies that hypological judgments do after all collapse into aretaic judgments. Although I reject this conclusion, it would still be significant, relying as it does on the central idea that, when it comes to judgments about moral responsibility, it is at bottom only responsibility *tout court* that matters; judgments about responsibility *for* something are essentially otiose.

---

[40] Or it may not. It may be that, had he, like Gray, had the opportunity to choose whether to embrace cruelty, Claggart, unlike Gray, would have freely chosen not to.

A final possible source of misunderstanding is my use of the term 'responsibility *tout court*'. It is intended to drive home the idea that we can be responsible without being responsible for anything. But, as I said earlier, it is not intended to suggest that responsibility can be nonrelational. In every case, Georg is responsible in virtue of some fact—the very same kind of fact in virtue of which George is responsible—even if in some of these cases (the ones involving situational luck) George is responsible for something while Georg is not. Nor should the term 'responsibility *tout court*' be thought to suggest that there is just one way in which an agent is responsible on any given occasion. On the contrary, since an indefinite number of counterfactuals about what one would do, if one were differently situated, can be true at once, one can be morally responsible *tout court*—both positively and negatively—to an indefinite number of degrees at once.[42] The view that I propose thus opens up the floodgates, as it were, when it comes to ascriptions of responsibility—of laudability as well as culpability.[43] The consequent profusion of ascriptions has profound practical implications, as I shall now briefly explain.

If, as I have urged, the truth of hypological judgments turns not just on what we actually freely do but, more deeply, on what we would counterfactually freely do, then the differences between individuals regarding the hypological judgments that are appropriate to them threaten to dissolve. (To what extent differences may still remain can only be a matter of speculation, depending on a number of issues that cannot be broached here.) For example, it may well be that not only Georg but most or even all of us would have freely acted as George did, were it not for some feature of our situation over which we lacked control. If so, we are, in virtue of this fact, deserving of the same reaction as George; if he deserves punishment, *then so do we*. By the same token, if George would have freely led the same sort of upright life as the lives that you and I lead, were it not for some feature of his situation over which he lacked control, then he is deserving of the same reaction as we; if we deserve to be rewarded (or at least not to be punished), *then so does he*. This casts considerable doubt on the propriety of our current practices, especially the practice of punishment. My point is *not* that no one can be culpable or deserving of punishment. Rather, it is that those whom we actually punish are likely to be no more deserving of punishment than many of those whom we do not punish and also likely to be as deserving of reward (or at least of nonpunishment) as many of those whom we reward (or refrain from punishing). Insofar as our current practices are based on judgments about what people deserve in light of the responsibility they bear, they radically distort the truth and are deeply discriminatory.

---

[41] I think this must be false. Virtues and vices are characteristics that are deeply entrenched and enduring. Being such that one would freely act in a certain way, given the opportunity, need not reflect or constitute any such characteristic.

[42] In "The Indeterminacy Paradox: Character Evaluations and Human Psychology" (unpublished), Peter B. M. Vranas argues for a similar view concerning aretaic judgments.

[43] This observation may take some of the sting out of the following complaint. Suppose that what accounted for Georg's not shooting Henrik was not a fortuitous sneeze but a sudden, unbidden crisis of conscience, over which he had no control and in the absence of which he would have freely shot Henrik. My view implies that Georg is just as culpable as George. Is this not too harsh? I do not think so. Note that my view also allows for the possibility that Georg is laudable, in virtue of being such that he would have (and did) freely *not* shot Henrik, given his crisis of conscience. (Perhaps, however, George is just as laudable as Georg in this regard, in that it might be that he too would have freely not shot Henry, had he been visited by an unbidden crisis of conscience.)

## VI

My argument here depends crucially on the claim that Georg would have freely killed Henrik, had certain features of the case cooperated. There are several reasons that may be given for denying this claim and, hence, for rejecting the argument.

One reason is this: we are never justified in making such a claim, at least when the antecedent is not satisfied. Even if it is true that Georg would have freely killed Henrik, if certain features of the case had cooperated, we cannot know this if such cooperation was, in fact, not forthcoming.

In response, let me grant that it is often likely to be true that we do not and cannot know such a thing, but I do not see why it should always be true. Suppose that Georg, stung by Henrik's insults, had plotted long and hard for revenge, that he had repeatedly voiced his desire to see Henrik dead, that the day had come to put his plan into action, that he had positioned himself in the appropriate place at the appropriate time, that he had raised his gun and was on the brink of shooting point-blank at an unprotected Henrik—and that he was interrupted by a sneeze at the crucial moment. And suppose that I was at his side every step of the way, witnessing all that took place.

Under such conditions, I think I would be justified in claiming that Georg would have freely killed Henrik, had he not sneezed when he did. But even if I am mistaken about this, my thesis here is unaffected; for this thesis depends, not on our knowing whether counterfactuals of this sort are true, but simply on their being true.

A second reason for doubting my thesis is this: such counterfactuals are never known to be true precisely because they never are true. But why think this? One answer would simply be that counterfactuals generally have no truth value. I shall not try to respond to this; I simply assume here that certain counterfactuals are true and that some account of them, perhaps along something like the well-known lines supplied by David Lewis,[44] is correct. Another answer is that, even though counterfactuals do generally have a truth value, counterfactuals of the sort at issue here, that have to do with free action in particular, are never true. Let me say something briefly about this.

It must be granted that it is possible that the sort of counterfactuals that are at issue here are never true. For example, even if we must accept the truth of the claim that, if certain features of the case had cooperated, it would have been the case that Georg either did or did not freely kill Henrik, it remains perfectly possible that it is not true either that, if there had been such cooperation, it would have been the case that Georg did freely kill Henrik, or that, if there had been such cooperation, it would have been the case that Georg did not freely kill Henrik (*ibid.*, pp. 16ff.). Perhaps this sort of thing sometimes happens; if it does, it will of course block the ascription of responsibility *tout court* to the agent. But although this may happen on occasion, why think that it is always actually the case?

It has been argued that, in fact, this must always be the case, if freedom is, as the libertarian conceives of it, incompatible with causal determinism. This argument is usually given in the context of a discussion whether divine foreknowledge is compatible with such freedom. Alvin Plantinga[45] has asserted that counterfactuals concerning such freedom can be, and often are, true (*ibid.*, pp. 173ff.). Adams and others[46] have argued to the contrary, on the grounds that counterfactuals require a necessary connection between antecedent and consequent, and that this conflicts with the libertarian view that freedom requires the absence of any such necessitation. This is a difficult matter which I cannot try to resolve here. Let me simply note the following. First, if freedom is compatible with causal determinism, then, even if Adams and others are correct, the argument that Georg is as responsible as George in all the various scenarios we have discussed is entirely unaffected. Second, if freedom is incompatible with causal determinism, then, even if Adams and others are correct (which is disputable), the argument, though admittedly affected, is not undercut nearly as drastically as it might at first appear. This is because, as Adams and others are ready to agree, even if it must be false that Georg would have freely killed Henrik, had he had the cooperation of certain features of the case which was, in fact, not forthcoming, it can nonetheless be true that Georg would probably have freely killed Henrik, had such cooperation been forthcoming, for as high a degree of probability as you like (short of certainty).[47] This may not satisfy the defender of the view that divine foreknowledge is compatible with libertarian freedom, insofar as God is supposed to be essentially infallible, but it suffices to show that Georg's blamelessness is by no means guaranteed. Suppose that there is a probability of .99 that Georg would have freely killed Henrik, had he not sneezed. Then one of two things follows: either Georg is 99% as responsible as George, or there is a 99% chance that Georg is as responsible as George. It is not clear to me which we should say, although I lean toward the latter.[48] In either case, Georg clearly cannot count on having a clean moral record just because he sneezed.

---

[44] *Counterfactuals* (Cambridge: Harvard, 1973). (Like Lewis (pp. 3-4), I am using the term 'counterfactual' broadly to cover even those cases in which the antecedent of such a statement is true.) A variation on the present misgiving is that counterfactuals are true only if a certain context of utterance is presupposed. I am not sure whether this is so; but, if it is, so be it. In that case, an ascription of responsibility will be accurate just in case it is grounded in a relevant counterfactual that is true in some context of utterance. Might it not happen that such a counterfactual is true in one context but not in another, so that the former context warrants an ascription of responsibility that the latter does not? I assume so. This would not affect the accuracy of the ascription associated with the former context. Might it not happen that the latter context warrants a *different* ascription of responsibility? Perhaps. But, again, this would not affect the accuracy of either ascription. Remember that multiple hypological judgments about one and the same agent in one and the same situation may all be accurate.

---

[45] *The Nature of Necessity* (New York: Oxford, 1974)

[46] Adams, "Middle Knowledge and the Problem of Evil," *American Philosophical Quarterly*, xiv, 2 (April 1977): 109-17; Peter van Inwagen, "Against Middle Knowledge," *Midwest Studies in Philosophy*, xxi (1997): 225-36; and Joshua Hoffman and Gary Rosenkrantz, *Exploring the Divine Attributes* (Malden, MA: Blackwell, 2002), chapter 6.

[47] Cf. Adams, "Middle Knowledge and the Problem of Evil," p. 111; van Inwagen, p. 231.

[48] Suppose that there are two mounds of some white substance on the table in front of you. There is a probability of 1 that the first is composed of sugar; there is a probability of .99 that the second is composed of sugar and a probability of .01 that it is composed of salt. We would say, not that the second is 99% as sweet as the first, but that there is a 99% chance that the second is as sweet as the first.

Of course, if Adams and others are right, then the question whether freedom is compatible with determinism recovers some of the significance that has traditionally been accorded to it. But if I am right that Georg is nonetheless probably as responsible as, or nearly as responsible as, George, this question certainly does not recover all of its significance.

A point related to the last one is this. Even if Adams and others are not right, might it not be that the truth of the claim that Georg would have freely killed Henrik depends on whether freedom is compatible with causal determinism? If so, would this not restore the significance of the question whether compatibilism is true?

This is a difficult matter. In principle, I do not see why there should not be considerable convergence between the judgments of compatibilists and incompatibilists regarding whether some agent would have freely done so-and-so under such-and-such circumstances.[49] To the extent that this is so, the question whether freedom is compatible with determinism is moot. To the extent that this is not so, however, I concede that the significance of this question is restored. But, even then, there is a strict limit to the restoration. As long as an incompatibilist is prepared to agree that some agent would, or would probably, have freely acted in some way in which he was caused not to act, and that this fact grounds the ascription of responsibility *tout court* to the agent, the question whether freedom is compatible with determinism simply does not recover all of the significance that has traditionally been accorded to it.

A final reason to be skeptical of the claim that Georg would have freely killed Henrik, had certain features of the case cooperated, has to do with those cases that concern constitutive luck in particular. It has been suggested that such luck is incoherent, inasmuch as it presupposes that one could have been a different person.[50] I deny this. What it presupposes is that one could have had different personal characteristics, and surely this is sometimes true. If Georg failed to kill Henrik simply because he was deaf (literally) to Henrik's insults, it seems clearly intelligible to say that he would, or would probably, have freely killed Henrik had he not been deaf. But what of other characteristics, such as being timid or thick skinned? Can we intelligibly say that Georg would have freely killed Henrik had he not had such characteristics as these? That depends on whether such characteristics are *essential* to Georg. I am inclined to think that they are not, but this is another difficult matter that I shall not try to resolve here.[51] Let me simply note that, if such characteristics are not essential to those who have them, then the truth of the relevant counterfactuals is unaffected. I concede, however, that, if any such characteristics are had essentially, then the relevant counterfactuals are indeed false.[52]

---

[49] See footnote 37 above, however.

[50] Cf. Daniel Statman, "Introduction," in Statman, ed., *Moral Luck* (Albany: SUNY, 1993), pp. 1-34, here p. 12; this volume contains reprintings of many papers, including those of Williams, Nagel, Rescher (revised), and my "Luck and Moral Responsibility." See also Rescher, p 14 in the original.

This concession is important. It means that the role that luck plays in the determination of moral responsibility may not be entirely eliminable, even if it is to be neutralized to the extent that I have argued for here. That is because, regardless of just which personal characteristics should be said to be essential to persons, it is presumably correct to say that some are. Consider any such characteristic that Georg may have. It is then necessarily false to say that Georg would have freely killed Henrik had he not had that characteristic. In such a case, Georg does get off the hook, even though he was, of course, not in control of whether he had the characteristic in question.[53] (Or rather, my argument does not suffice to keep him on the hook. The general thesis that one cannot escape responsibility through luck should incline one to think that Georg can be responsible even in a case in which he fails to act freely due to some characteristic that is essential to him, since the possession of any such characteristic is a matter of luck.) But this concession is not enough to resurrect the significance traditionally accorded to the questions whether freedom requires alternate possibilities and whether freedom is compatible with causal determinism. I am prepared to agree that the *capacity* to act freely is essential to whatever has it. (Thus it must be false to say, for instance, that Georg's *gun* would have freely killed Henrik, had *it* had the capacity to do so.) But it is surely not correct to say that one's being such that one has (or lacks) alternate

---

[51] It should be noted that my position here is consistent with the claim made earlier that Melville's character, Claggart, is uncontrollably evil. A characteristic may be uncontrollable by a person and yet not be essential to that person. This is clearly true when the characteristic in question is not that person's but someone else's; it is no less true when the characteristic is the person's own.

[52] Lewis himself declares (*op. cit.*, pp. 16, 24-26) that counterfactuals with impossible antecedents are vacuously true, which strikes me as inadvisable. But he grants (*op cit.*, p. 25) that his reasons are "less than decisive" and provides an alternative account according to which such counterfactuals are necessarily false.

[53] Cf. Greco, pp. 94-95.

# VALUE-PLURALIST EGALITARIANISM*

For over two decades now, egalitarian analytical philosophers have puzzled over what dimension is to be equalized in the pursuit of equality.[1] For example, Amartya Sen[2] has advocated "basic capability equality," and in the process has argued both against equalizing marginal utility and against adopting a Rawlsian approach. Because, primarily, of the problem posed by expensive tastes, Ronald Dworkin[3] has famously argued against equalizing welfare, and has powerfully advocated equalizing resources instead. Richard Arneson[4] has agreed that "equality of welfare is a poor ideal" (*ibid.*, p. 83), but he has also claimed that Dworkin has not, in fact, argued for equality of resources. Rather, he has argued for what Arneson labels "equal opportunity for resources" (*ibid.*, p. 93, note 12). But contrasting equal opportunity for resources with equal welfare is not to compare like with like. And if one compares equal opportunity for resources with equal opportunity for welfare—which would be a case of comparing like with like—then equal opportunity for resources is preferable, in Arneson's view. But G. A. Cohen[5] prefers what he calls "*equal access to advantage*, where 'advantage' is understood to include, but to be wider than, welfare" (*ibid.*, p. 907). Why, then, does Cohen prefer the term 'access' to the term 'opportunity'? Because

> [w]e would not normally regard meager personal capacity as detracting from opportunity. Your opportunities are the same whether you are strong and clever or weak and stupid: if you are weak and stupid, you may not use them well—but that implies that you have them. But shortfalls on the side of personal capacity nevertheless engage egalitar-

---

* I am grateful to John Baker and G. A. Cohen for their comments on an early draft of this discussion, and to Dick Arneson, Matthew Clayton, David Donaldson, and Andrew Williams for their comments on a later version. I am also indebted to David Boonin-Vail and, especially, Michael Tooley for forcing me to consider nonparabolic indifference curves.

[1] I refrain here from discussing Larry Temkin's views on the differing ways in which a situation can be regarded as equal or unequal. For a critique of Temkin, see my "Simplifying 'Inequality'," *Philosophy and Public Affairs*, xxx, 1 (2001): 88-100.

[2] "Equality of What?" in *Choice, Welfare and Measurement* (New York: Blackwell, 1982), pp. 353-69.

[3] "What Is Equality? Part 1: Equality of Welfare," *Philosophy and Public Affairs*, x, 3 (1981): 185-246 and "What Is Equality? Part 2: Equality of Resources," *Philosophy and Public Affairs*, x, 4 (1981): 283-345.

[4] "Equality and Equal Opportunity for Welfare," *Philosophical Studies*, LVI (1989): 77-93.

[5] "On the Currency of Egalitarian Justice," *Ethics*, XCIX (July 1989): 906-44.

0022-362X/02/9911/577-99

> ian concern, and they do so because they detract from access to valuable things, even if they do not diminish the opportunity to get them (*ibid.*, pp. 916-17).

And Cohen prefers 'advantage' to 'welfare' because both "resource deficiencies and welfare deficiencies are distinct types of disadvantage" (*ibid.*, pp. 920-21), and both trouble the egalitarian.

## I. A PLURALITY OF VALUES

In short, a number of famous attempts have been made at identifying the dimension that egalitarians should seek to equalize: the meeting of basic capabilities, resources, opportunity for welfare, and access to advantage have all been put forward as the much-needed metric. But when one considers the various core arguments that have been leveled against each conception of equality proposed—arguments that have seemed persuasive to many egalitarian analytical philosophers—then it would appear that, at the end of the day, each conception is unattractive because it violates some or other value prioritized in one of the other conceptions of equality.[6]

Consider Sen's primary objection against what he calls "utilitarian equality": an imaginary pleasure wizard, who was extremely efficient at converting resources into welfare, would enjoy a highly disproportionate degree of welfare if the goal were to maximize utility. This unpalatable result appears to be demanded by utilitarianism because of the theory's "single-minded concern with maximizing the utility sum" (*op. cit.*, p. 357). Certainly, the variety of utilitarianism Sen is here criticizing could lead to a highly unequal distribution of welfare. But it also seems to lead, through its unqualified pursuit of increased welfare, to a highly unequal distribution of resources. And the latter may be what is doing a great deal of the work in persuading an egalitarian audience that "utilitarian equality" is unacceptable. On the other hand, Sen rejects "Rawlsian equality" because the distribution of primary social goods can leave the disabled severely disadvantaged (*op. cit.*, p. 366). Given that this objection is presented before Sen introduces his notion of "basic capabilities equality," it may well do most of its work by appealing to the egalitarian's odium toward significant differences in welfare.

---

[6] A word of caution: in what follows I seek to clarify the structure of egalitarian thinking. The question of whether or not one ought to be an egalitarian is a wholly separate issue. Furthermore, were universal agreement to be reached that a significant reduction in average utility, say, would be a very bad thing, then a committed egalitarian would surely seek to incorporate average utility as a value within her egalitarianism before regarding utilitarian considerations as simply demanding the abandonment of egalitarian pursuits.

Or consider Dworkin,[7] who is, without doubt, the principal critic of those who would seek to equalize well-being. His rejection of welfarist egalitarianism does not prevent him from recognizing the "immediate attraction of equality of welfare" (*ibid.*, p. 14), nor from acknowledging that its claims "are both familiar and apparent" (*ibid.*, p. 13), nor from noting that "[t]here is an immediate appeal in the idea that insofar as equality is important, it must be equality of welfare that counts" (*ibid.*, p. 14). Its immediate appeal is, of course, precisely what motivates Dworkin to devote so much time and space to a rejection of equality of welfare. Yet consider one of his arguments—a seemingly knock-down one—against the pursuit of such equality. Egalitarians are asked to

> [s]uppose that the welfare (on any interpretation) of an entirely paralyzed but conscious person is vastly less than the welfare of anyone else in the community, that putting more and more money at his disposal would steadily increase his welfare but only for very small amounts, and that if he had at his disposal all the resources beyond those needed simply to keep the others alive he would still have vastly less welfare than they.[8]

If equalizing welfare were the egalitarian's goal, then she would have to demand a radical transfer of resources into the hands of such an unfortunate. But this seems more than a little excessive. And given that such a severely disabled person would, in effect, command most of the society's resources, it would also strike many as highly inegalitarian.[9] Moreover, Dworkin's thought experiment is not as wild as it might initially appear. For, clearly, some individuals *are* so unfortunate that they would require a massive proportion of a community's resources simply to raise their level of welfare slightly, while everyone else would, thereby, be reduced to poverty. One need merely think of those who suffer from what is currently an incurable disease. Research into cures for diseases that afflict one in a million could easily take up a vast proportion of any society's resources if it were serious about seeking to equalize the welfare of the one in a million with that of the rest.

Dworkin's argument is certainly a powerful one. But its power seems to rest on the violation of at least *two* values, and not merely on the disvalue many place on a highly unequal distribution of resources. As in the case of Sen's pleasure wizard, attending to the needs of the entirely paralyzed person in Dworkin's example does require an unacceptably unequal distribution of resources. But also, insofar as egalitarians value a high average welfare, any attempt to equalize the welfare of so extreme an unfortunate with that of the others would result in an unacceptably low average utility level. Indeed, when Dworkin mentions a possible outcome being one where the unfortunate is still at a lower level of welfare than others, while having "at his disposal all the resources beyond those needed simply to keep the others alive,"[10] it may well be the low average utility which makes such an attempt at attaining equality of welfare appear so unacceptable—perhaps more so than either the ultimate impossibility of achieving it, or the massively unequal distribution of resources that would be involved in the unsuccessful attempt.

---

[7] *Sovereign Virtue: The Theory and Practice of Equality* (Cambridge: Harvard, 2000).

[8] "What Is Equality? Part 1: Equality of Welfare," p. 242.

[9] Interestingly, this argument seems to tell equally against equalizing basic capabilities. Hence, it appears to be just as telling against Elizabeth S. Anderson's development of Sen's "capability egalitarianism"—"What Is the Point of Equality?" *Ethics*, cix, 2 (1999): 287-337.

Now consider the various arguments that have been deployed by egalitarians against expensive tastes. Many appear to trade on the egalitarian's abhorrence of extremely unequal distributions of resources. But expensive tastes also reduce the average utility in a society. And this might well provide a further reason why egalitarians often feel that their acquisition ought to be discouraged.[11] And it seems clear upon reflection that, initial appearances to the contrary, egalitarians are, on the whole, extremely concerned about the level of average welfare. For if it were of no concern to them, why should so many egalitarians feel (as they clearly do) the need to refute the often heard charge that equality would greatly reduce the average level of utility?[12]

---

[10] "What Is Equality? Part 1: Equality of Welfare," p. 242.

[11] Dworkin, however, thinks that this does not explain our discomfort with compensating expensive tastes: "Expensive tastes are embarrassing for the theory that equality means equality of welfare precisely because we believe that equality, considered in itself and apart from questions of efficiency, condemns rather than recommends compensating for deliberately cultivated expensive tastes" (*ibid.*, p. 235). But if it is the case that many egalitarians also value striving for the highest attainable level of average utility, then the cultivation of expensive tastes, which would ordinarily result in deflating the average utility level, could well be what is doing a great deal of the work in Dworkin's objections to equality of welfare. And it may be that it is only because he is already committed to equality of resources that he dismisses widespread egalitarian concerns about reductions in average utility.

[12] It is, perhaps, worth noting that a follower of Rawls could attempt to defend the difference principle by arguing that egalitarians surely care about each person's welfare, and therefore should not object to inequalities which would increase everyone's welfare. But such an argument, if it is to have any purchase on egalitarians, must rely upon their concern for something like the average level of utility.

Focusing upon the level of average welfare seems prey to an obvious objection, however. Derek Parfit[13] asks us to

> [s]uppose...that the people in some community could all be either (1) equally well off, or (2) equally badly off. The Principle of Equality does not tell us that (2) would be worse. This principle is about the badness of inequality; and, though it would be clearly worse if everyone were equally worse off, our ground for thinking this cannot be egalitarian (*ibid.*, p. 4).

Thus, focusing upon the level of average welfare appears at first glance to be irrelevant to any discussion of egalitarianism. But given that few, if any, self-proclaimed egalitarians would be indifferent between (1) and (2), then Parfit would seem to be on firm ground when he observes that "equality cannot plausibly be our only value. If we are Egalitarians, we must hold some more complicated view" (*ibid.*, p. 22).

What all this suggests is that egalitarians, insofar as they perceive some force in all the arguments for and against basic capabilities equality, equality of welfare or resources, equal opportunity for welfare, or equal access to advantage, in fact subscribe to *a plurality of values*.[14]

### II. EQUALITY OF RESOURCES *AND* EQUALITY OF WELFARE?

The subscription to a plurality of values is perhaps most obvious in the case of Cohen, who, as we have noted, claims that both "resource deficiencies and welfare deficiencies are distinct types of disadvantage" (*op. cit.*, pp. 920-21) and that both matter to the egalitarian. In order to convince us of this, Cohen points out that egalitarians would want to give a wheelchair to someone who cannot walk, even if that person were already extremely happy. This seems to show that equality of welfare will not provide the norm egalitarians seek. But as Cohen points out: "it also defeats equality of opportunity for welfare, since the response to disability is shaped by something other than the different costs in lost opportunity for welfare which disability causes in different people" (*op. cit.*, p. 918). The egalitarian response must be that those who cannot walk need wheelchairs in order to be resourced adequately.

But equality of resources would not seem to provide the sought-after norm either. For Cohen asks us to imagine someone who can move his arms more easily than most other persons can, but that any arm movement causes him great pain later. If there is an expensive medicine that removes the pain, egalitarians would surely want the medicine to be provided. But as Cohen points out: "providing the medicine cannot be represented as compensating for a resource incapacity" (*op. cit.*, p. 919). For this particular person, while it is true that he feels pain after moving his arms, can nevertheless move them more easily than most people are able to. So, in Cohen's view, there must be more than one dimension to disadvantage.

---

[13] *Equality or Priority? The Lindley Lecture* (Lawrence: Kansas UP, 1995).

[14] Thus, it would seem that, to employ Parfit's terminology, most egalitarians are not "pure egalitarians" but "pluralist egalitarians" (*ibid.*, p. 5).

This would also appear to constitute a decisive objection to Dworkin's equality of resources.[15] But consider his recent response:

> Almost everyone would agree that a decent life, whatever its other features, is one that is free from serious and enduring physical or mental pain or discomfort, and having a physical or mental infirmity or condition that makes pain or depression or discomfort inescapable without expensive medicine or clothing is therefore an evident and straightforward handicap. Someone with such an infirmity did not choose it; he would cure it if he could, and none of his beliefs, judgments, convictions, or commitments would argue against such a cure. A pain-producing infirmity is a canonical example of a lack of personal resources for which equality of resources would, in principle, provide compensation.[16]

And as he adds:

> Cohen apparently believes that this obvious response to his claim that equality of resources would not compensate for pain appeals to well-being or welfare in a way that compromises my arguments against equality of welfare. But I believe that he misunderstands [my] arguments.... I did not mean to bar political philosophy from considering people's well-being in any way—that would be mad—but only to deny equality of welfare as a metric of justice. When we consider which features of a person's physical and mental constitution we should count as handicaps eligible for compensation, we must of course consider the consequences of these features for his well-being: mental illness that produces serious and enduring depression is a handicap for that reason. It does not follow that we must accept that any feature of his personality that lessens his happiness, including his tastes and convictions, is a handicap.... In any case, providing compensation for the unfortunate

---

[15] It might be objected that our intuitive responses are insufficient to constitute a decisive objection. For example, if one believed that seeking reflective equilibrium were the appropriate method to adopt in moral philosophy, then certain intuitive reactions to possible situations might well need to be rejected. But Cohen's objection to Dworkin's dismissal of welfarism seems to carry weight even upon considered reflection. And this would appear to be because it appeals to deep values.

[16] *Sovereign Virtue*, p. 297.

> people Cohen imagines does not assume that we take equality of welfare as a goal. If the community gives someone money for medicine to relieve pain, it does so not in order to make his welfare or well-being equal to anyone else's, but because his physical constitution handicaps his ability to lead the life he wishes to lead (*ibid.*, pp. 490-91, note 11).

But while pain which is incapacitating might be the only pain that concerns Dworkin's conception of justice, I doubt that the majority of egalitarians would share only such a severely restricted concern. Moreover, and most importantly, it is hardly surprising that Dworkin should feel confident in reducing "equality" to "equality of resources," for his notion of "equality of resources" appears to incorporate what almost everyone else would regard as a major aspect of welfare, namely, freedom from pain. And it is unsurprising, therefore, that Dworkin's position should have been so misunderstood, and not just by Cohen.

Clearly, of the major egalitarian theorists Cohen seems the most ready to trade on a plurality of values. But it is not clear that all of the values doing the work in supporting his notion of equal access to advantage are straightforwardly what he would regard as egalitarian ones. Cohen is explicit that other values might defeat egalitarian concerns (*op. cit.*, p. 906). But, prima facie at least, it seems odd that, in order to avoid having to compensate for expensive tastes that have been intentionally cultivated, Cohen appears to place at the heart of his conception of equality a value ordinarily thought to contradict egalitarianism, namely, choice.

Cohen might claim, however, that in putting choice in the foreground, he is merely recognizing the unfairness that would result from the cautious having to subsidize the mistakes of the risky, and that such unfairness might appear unjust. But some might regard as unfair the costs that the risky could end up bearing. Indeed, some might object, is Cohen's position not rather too cavalier with the costs that certain choices might incur? Of course, there are many everyday choices that most of us would want individuals to pay the cost or reap the benefits of. On the other hand, as John Baker[17] implores: "there are big choices in life which are full of uncertainty, or depend on circumstances, upbringing, emotion, and influence. Choices of education, of job, of spouse, of residence, even of lifestyle—if these choices turn out badly, is it really right to blame the people who make them, and to insist that they bear the consequences? Why not give them a second chance" (*ibid.*, p. 67)? And as he adds: "A policy of unrelieved suffering for past mistakes can't plausibly be described as a way of sharing equally the burdens of choice" (*ibid.*).[18] This suggests that there are egalitarian reasons for why there ought to be limits to the costs (and perhaps the benefits) accruing from individual choices.

---

[17] *Arguing for Equality* (New York: Verso, 1987).

But then, Cohen's prioritization of choice would not seem to be fully justified simply on the basis of fairness or justice. What value, then, would justify prioritizing choice? If anything, one would expect it to be the value of freedom or autonomy. Moreover, if we value equal access to advantage, and if we are to discount the outcomes of choice, it is most likely because we value equal autonomy. Indeed, the most plausible reason why choice is of such significance is surely because we cannot have autonomy without having choice. And a failure to factor in choice is tantamount to a failure to respect autonomy. But if individuals are to be genuinely autonomous, then, it might be assumed, they will have to accept responsibility and pay (at least some of) the costs of their choices. Hence, it would seem that it is only because of a covert emphasis on the value of freedom or autonomy that Cohen is unconcerned by certain unequal distributions of resources and welfare.

In a word, it would appear that egalitarians subscribe to a far more complex set of ideals than is often explicitly acknowledged. For it would seem from the plethora of arguments offered on behalf of the various construals of egalitarianism that egalitarians subscribe to a plurality of values when making their respective cases, and, less contentiously, that *an egalitarian audience subscribes to a plurality of values in responding favorably to the various arguments proffered.* This would suggest that the reason why the search for *the* metric of equality has failed is because theorists of equality have sought to identify one dimension, when the values of egalitarians are, rather, multidimensional. And this further suggests that the reason why unidimensional conceptions of equality fail is because, in emphasizing only one value, they have implications that offend the other values to which egalitarians frequently subscribe. And the reason for this is that the several values that egalitarians usually hold are not compossibly realizable.

## III. PLURAL VALUES AND EGALITARIANISM

Many self-proclaimed egalitarians are well aware that the pursuit of some of our values can be defeated by others. For example, Dworkin writes:

> It might be that wealth disparities would be so greatly reduced by the features of the economy we have already described that we would be

---

[18] Also see Anderson, pp. 295-96.

> much less troubled than we might suspect in advance by the wealth inequalities that would remain. Indeed it might be that the costs in overall efficiency of even those features would be so great that those who are prepared to compromise equality of resources either for general utility or in service of some strategy of making the worst-off as well-off as possible, would argue that even that much equality would be condemned by their more embracing conception of justice.[19]

Alternatively, take Cohen, who remarks:

> An *equalisandum* claim specifies that which ought to be equalized, what, that is, people should be rendered equal in. An unqualified or *strong equalisandum* claim, which is the sort that an uncompromising egalitarian asserts, says that people should be as equal as possible in the dimension it specifies. A qualified or *weak equalisandum* claim says that they should be as equal as possible in some dimension but subject to whatever limitations need to be imposed in deference to other values... (*op. cit.*, p. 908).

And Cohen seeks only to defend a weak *equalisandum* claim.

But what is, perhaps, even more interesting is the following. Cohen, in a humble mood, indicates that he considers his own egalitarian proposal to be somewhat crude, expressing his hope that "it will be superseded" because of

> the unlovely heterogeneity of the components of the vector of advantage. One hopes that there is a currency more fundamental than either resources or welfare in which the various egalitarian responses which motivated my proposal can be expressed. But I certainly have not discovered it, so, at least for now, I stay with the appearances, which contradict welfare, resources, and opportunity for welfare readings of the egalitarian demand, and which point, in the first instance, to the theory (or semi-theory: it is perhaps too close to the intuitive phenomena to merit the name "theory") I have affirmed (*op. cit.*, p. 921).

And, perhaps, most telling of all, while commenting on his treatment of Dworkin's famous example of Jude,[20] Cohen exclaims: "I do not feel comfortable about this victory, since, in achieving it, I exploit to the hilt a feature of my theory which I regard as suspect: the heterogeneity of its conception of advantage" (*op. cit.*, p. 925). While Cohen[21] values both equality of welfare and equality of resources, he explicitly expresses dissatisfaction with his "semi-theory" because it is heterogeneous, combining elements from both equality of welfare and equality of resources. I would argue, however, that this is not a weakness in Cohen's approach but, rather, its strength. Nevertheless, Cohen hopes that, in the final analysis, there will be revealed to be one, and only one, egalitarian legal tender, and thus gives the impression that one should persist in the search for a single metric to act as the universal currency. But given the seeming plurality of values shared by egalitarians, surely what we require, instead, is something more akin to an egalitarian *bureau de change*. Allow me to suggest what one might look like.

---

[19] "What Is Equality? Part 2: Equality of Resources," p. 331.

[20] Dworkin asks us to "[i]magine...a society newly dedicated to equality of enjoyment in which, when resources are redistributed in order to achieve equality of enjoyment, Jude has far less money than anyone else because his wants are so simple and so inexpensively satisfied. But one day (perhaps after reading Hemingway) he decides that his life, for all its richness and enjoyment, is a life of less overall success than it might be, and proposes to cultivate a new taste for some challenging sport, such as bullfighting. Suppose that after he does so he finds himself seriously frustrated by his lack of funds with which, for example, to travel to Spain, and asks for more funds through a further redistribution after which he would still, as things fall out, have less than anyone else. Do we now have any grounds for saying that he

## IV. LEXICAL ORDERING OR VALUE TRADE-OFFS?

Elsewhere,[22] I argue that each of the prevailing moral theories—Aristotelianism, Kantianism, and utilitarianism (in both its total and average variants)—is flawed because it attempts to maximize one value, and in so doing ultimately reveals its inadequacy by flouting the other values we hold. As I also argue, however, there is an exception: John Rawls's[23] theory.

Rawls acknowledges a plurality of values, but lexically orders them. The ostensible ordering is such that the value we place on freedom is set above the value we place on equal opportunity, which is itself set above the value we place on an increase in income. That, at least, is how Rawls's two principles of justice appear to be presented. But his theory is, in actual fact, far messier than this description suggests.

One reason for Rawls's lexical ordering is that there are restrictions on liberty which, he presumes, no rational person would accept merely for an increase in income. As Rawls is well aware (*ibid.*, p. 542),

---

is undeserving of the increase, when we know that if it is denied he will have both less funds and less enjoyment than anyone else? I doubt anyone will want to say this... Jude asks only that something closer to an equal share be put at the disposal of his [life]. We need the idea of fair shares (in this particular case the idea of an equal share of resources) in order to express the force of this..."—"What Is Equality? Part 1: Equality of Welfare," p. 239.

[21] *If You're an Egalitarian, How Come You're So Rich?* (Cambridge: Harvard, 2000), p. 177, where Cohen, in the attached endnote, immediately refers the reader back to his "On the Currency of Egalitarian Justice."

[22] See my "Moral Theory and Global Population," *Proceedings of the Aristotelian Society*, xcix, 3 (1999): 289-313.

[23] *A Theory of Justice* (Cambridge: Harvard, 1971).

however, it is undeniable that some levels of income are so low that it would be quite rational to trade liberty for a gain in income. Thus, it is only to societies that have attained a certain level of wealth that Rawls's lexically ordered principles are supposed to apply. But this means, in effect, that a certain level of income is, in fact, lexically prior to political liberties,[24] which are lexically prior to fair equality of opportunity, which in turn is lexically prior to an increase in income! But then, there is a sense in which political liberties are both lexically prior and lexically subordinate to income and wealth.

Once it is noticed that Rawls's lexical ordering is messier than it first appears, it becomes less clear that he has given us a correct description of the relationship between our various values. Remarkably, Rawls himself acknowledges as much. As he writes: "While it seems clear that, in general, a lexical order cannot be strictly correct, it may be an illuminating approximation under certain special though significant conditions."[25] But if it is only illuminating under certain conditions, can we, perhaps, provide a more general formulation of how our various values relate to one another?

One possible way of doing so is suggested in the early work of Brian Barry.[26] According to Barry: "The problem of someone making an evaluation can...be regarded as the problem of deciding what mixture of principles more or less implemented out of all the mixtures which are available would be, in his own opinion, best" (*ibid.*, p. 6). But how might we best view such a mixture? Barry posits that, if we take any two general principles, "for each person who evaluates in terms of these principles we can draw up a set of indifference curves showing along each line different combinations of the two between which he [or she] would be indifferent" (*ibid.*, p. 5). Taking Barry's own examples, as it is possible to trade equity for efficiency, and vice versa, and as one can be indifferent between a gain in efficiency and a loss in equity, then any such trade-offs we deem acceptable can be plotted on a graph.

Perhaps, then, the relationship between our several values is better modeled by indifference curves than by Rawls's lexical ordering. Consider Figure 1, which models how an individual might regard certain increases in income as substitutable for specific losses in liberty, and vice versa. Income is measured on the *X*-axis and liberty on the *Y*-axis. Any point on *aa'* is viewed as of equal value to any other

---

[24] Rawls has come to admit that this is a lexical ordering—*Political Liberalism* (New York: Columbia, 1996), p. 7.

[25] *A Theory of Justice*, p. 45.

[26] *Political Argument* (New York: Humanities, 1965).

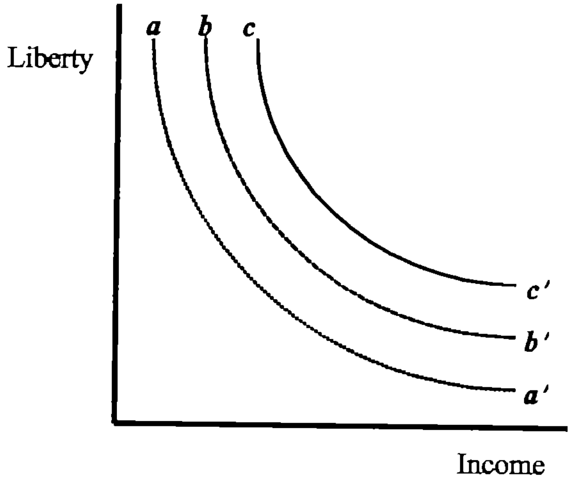


Figure 1: An income/liberty indifference map.

point on that curve, but is regarded as considerably less preferable to any point on *bb'*, which in turn is regarded as considerably less preferable to any point on *cc'*.

Now, were any point on *cc'* available, point *a* would then represent a situation where, though there is a great deal of liberty, there is so little income that the situation would appear unacceptable, for it falls on a far lower indifference curve. Moreover, given the availability of some point on *cc'*, then point *a'* would represent a situation where, although the income is extremely high, there is so little liberty that this situation, too, would appear equally unacceptable. Rawls's ordering, where income seems to be both lexically prior and lexically subordinate to political liberties, could be construed as an attempt to avoid the situations represented by points *a* and *a'*.

But if the truth in Rawls's theory is simply that some income levels are so low and some liberties so few that, with alternative distributions being available, no gain in the other could compensate, then, in fact, neither liberty nor income would be *lexically* prior. Furthermore, the possible trade-offs along *cc'* mapped in Figure 1 rule out points *a* and *a'* no less stringently than Rawls's ordering does. Point *a*, in falling well beneath an available point on *cc'*, represents a situation in which no increase in liberty would compensate for the poverty endured, while point *a'*, in also falling beneath the available point on *cc'*, represents a situation in which no increase in income would compensate for the lack of liberty. In short, the difference between such indifference curves and Rawls's lexical ordering mainly concerns the possible trade-offs between liberty and income which fall along the indifference curves *aa'*, *bb'*, and *cc'*.

Is Rawls right, then, to dismiss such trade-offs? It is not clear that the answer is "yes." Trade-offs between liberty and income are often made in everyday life. In their quest for higher income, employees frequently accept restrictions on what they can say and do. For pay, they sometimes "sell" their liberty to discuss their employer's secrets (whether industrial or state),[27] to take one example. In which case, indifference curves may better represent the relations between our several values than Rawls's lexical ordering does. And as Rawls himself concedes: "there is nothing intrinsically irrational about this intuitionist doctrine. Indeed, it may be true."[28]

## V. TRADING EQUALITY OF WELFARE FOR EQUALITY OF RESOURCES

Earlier, I argued that egalitarians appear to display an adherence to a plurality of values. In particular, given their receptivity to the various arguments deployed by the analytical philosophers I have mentioned, egalitarians would seem to value *both* equality of welfare and equality of resources. But because of the different capacities different individuals have for converting resources into welfare, an absolute equality of welfare would entail substantial inequalities in the distribution of resources, and vice versa. But if egalitarians value both equality of welfare and equality of resources, and if they cannot maximize both simultaneously, then it seems that they will have to consider the various trade-offs that would need to be made in the pursuit of both values. Such possible trade-offs are likely to take the parabolic form represented in Figure 2, where the valuing of both equality of welfare and equality of resources by a representative egalitarian would probably lead her to require progressively greater increases in one in order to compensate for any further losses in the other.

But let us forgo relying on this assumption. Imagine, instead, that egalitarians would trade increases in equality of welfare for decreasing equality of resources, and vice versa, in constant proportions, as represented in Figure 3. In Figure 3 (as well as in Figure 2), all points on *aa'* are viewed as of equal value, but are regarded as considerably less preferable to any point on *bb'*, all points on *bb'* being regarded as

---

[27] Many people seem willing to exclude certain "state secrets" from freedom of information and discussion, especially if the apparent purpose of those secrets is to protect their present way of life.

[28] Rawls, *A Theory of Justice*, p. 39. It is worth noting that nothing I am arguing presupposes a mysterious faculty that "intuits" moral properties. And neither does Rawls's idiosyncratic use of "intuitionism" have any such implication.

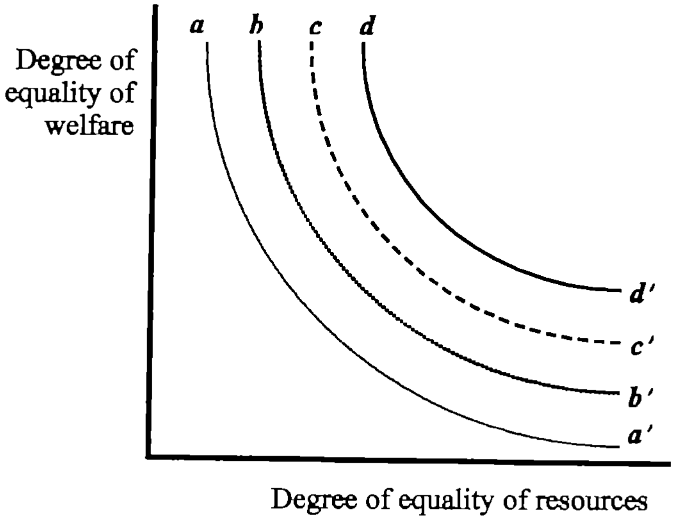

Degree of equality of welfare
a
b
c
d
d'
c'
b'
a'
Degree of equality of resources

Figure 2: An equality of resources/equality of welfare indifference map.

considerably less preferable to any point on *cc'*, and all points on *cc'* being regarded as considerably less preferable to any point on *dd'*.

But while such an indifference map might represent the trade-offs an egalitarian might make, we do not, as yet, know which combina-

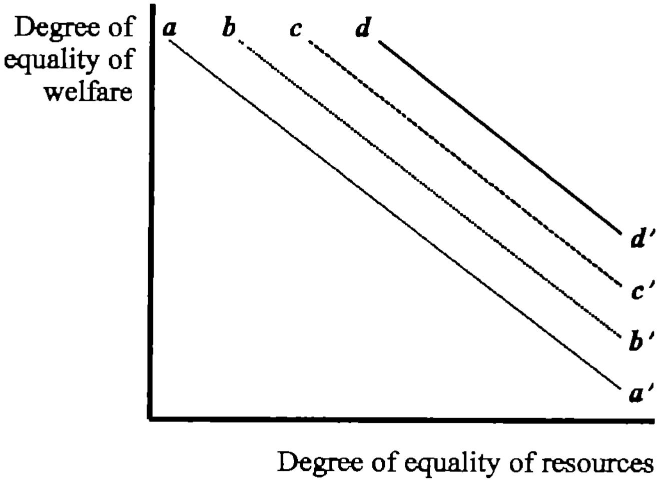

Degree of equality of welfare
a
b
c
d
d'
c'
b'
a'
Degree of equality of resources

Figure 3: An alternative equality of resources/equality of welfare indifference map.

tions of welfare distribution and resource distribution would actually be available or feasible. If there were a completely equal distribution of resources, then there would be considerable inequalities in welfare (given the different capacities individuals possess for deriving welfare from resources). But it is not difficult to imagine situations where introducing a small inequality in resources would lead to a considerable increase in the equality of welfare. For example, were equality of resources to obtain, if half the population were suffering from hangovers, then distributing a relatively cheap painkiller (such as aspirin) to them, and to them alone, would greatly increase the equality of welfare while creating only a very minor inequality in resource distribution. So, when equality of resources is measured on the *X*-axis and equality of welfare on the *Y*-axis, the slope at this point would be very steep.

On the other hand, it is not difficult to imagine situations where a slight increase in equality of welfare would require a massive inequality in the distribution of resources. For example, as I mentioned earlier, some people suffer from extremely rare illnesses that are, at present, incurable. Were individual welfare levels fairly equal, then in order to equalize further everyone's welfare, an astronomically high proportion of a society's resources might have to be devoted to the search for cures to illnesses that affect one person in a million. And that would be tantamount to distributing an exceedingly disproportionate amount of that society's resources to such people. So, when equality of resources is measured on the *X*-axis and equality of welfare on the *Y*-axis, the slope at this point would be very gradual.

This suggests that the combinations of resource distribution and welfare distribution which are possible would, when viewed from the origin, be represented by a concave curve. So, let Figure 4 represent the various combinations of welfare and resource distributions that could, in fact, be actualized. In other words, the available combinations all fall along $vv'$.

Let us now superimpose the availability curve (or feasibility curve) $vv'$ on the indifference curves mapped out in Figure 3. This generates Figure 5. And what Figure 5 clearly shows is that, starting from point $t$, where the availability curve $vv'$ touches the highest indifference curve ($dd'$), the closer one moves toward either axis—toward either point $v$ or point $v'$—the lower (or closer to the origin) is the indifference curve that $vv'$ bisects. In other words, from the perspective afforded by point $t$, significant moves toward an increase in the equality of welfare or in the equality of resources will become progressively less desirable. Indeed, from the standpoint of $t$, which is a combination of welfare and resource distributions which is actually

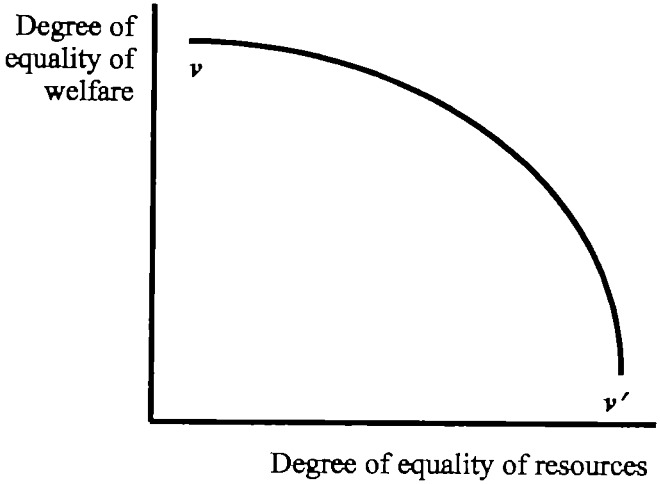

Degree of equality of welfare
v
v′
Degree of equality of resources

Figure 4: An equality of resources/equality of welfare availability curve.

feasible, the maximizing of equality of welfare—represented by point $v$—or the maximizing of equality of resources—represented by point $v'$—will appear positively undesirable, given that both points fall on a very low indifference curve ($aa'$). Moreover, if the indifference curves were to take the parabolic shape of those in Figure 2, the progressive

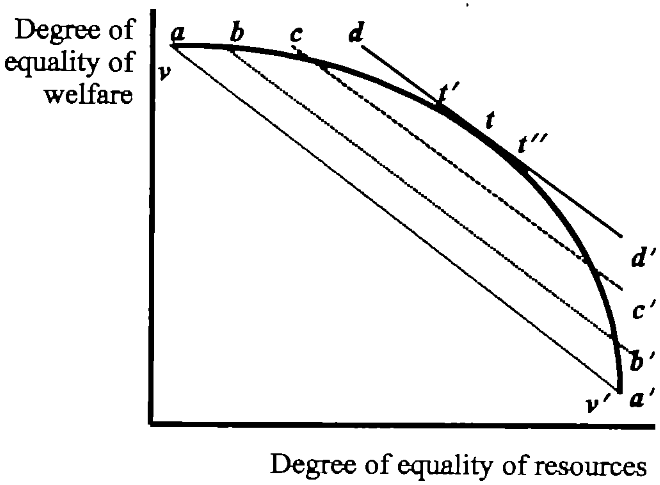

Degree of equality of welfare
a
b
c
d
v
t′
t
t″
d′
c′
b′
v′
a′
Degree of equality of resources

Figure 5: An equality of resources/equality of welfare indifference map and actuality curve.

undesirability of further increases in either equality of welfare or equality of resources would be even more pronounced.

In addition, Figure 5 also shows that, as points $t'$ and $t''$ (falling on the availability curve $vv'$) barely fall below $dd'$ (the highest indifference curve), then there is not a great deal from which to choose between points $t$, $t'$, and $t''$.[29]

What all this suggests, surprising as it might at first seem, is that value-pluralist egalitarianism—the variety of egalitarianism that seems to be, often unconsciously, presupposed in egalitarian debates—will not be greatly concerned with enforcing an exact distribution. But it will be especially concerned both with a significantly unequal distribution of welfare, even when it involves an equal distribution of resources (such as at point $v'$ in Figure 5), and with a significantly unequal distribution of resources, even when it involves an equal distribution of welfare (such as at point $v$ in Figure 5).

Earlier, we noted that Dworkin entertains the possibility of a "more embracing conception of justice."[30] Can we go beyond simply comparing equality of welfare and equality of resources? Could indifference maps, plotting our various values and how they relate, give some indication of what a more embracing conception might be like?

## VI. MULTIDIMENSIONAL EGALITARIANISM

Clearly, a more embracing conception of justice than Dworkin's equality of resources would require the employment of more than two values.[31] The potential trade-offs involved in one such conception are represented diagramatically in Figure 6, where the axes measure degree of equality of welfare, degree of equality of resources, and level of average utility. Of course, given the seeming plurality of our values, an even more embracing conception of justice would require an axis measuring, perhaps, the level of autonomy, perhaps one measuring (inversely) the extent of rights violations, and so on. Mathematicians are not unfamiliar with multidimensionality. But given the difficulty in *drawing* multidimensional indifference maps, I shall confine my discussion to the values measured along the three axes of Figure 6.[32]

---

[29] The more parabolic the indifference curves, however, the closer would points $t$, $t'$, and $t''$ need to be—if, that is, one is to be relatively indifferent between them.

[30] "What Is Equality? Part 2: Equality of Resources," p. 331.

[31] To be fair to Dworkin, however, it should be noted that he incorporates certain other values within his conception of equality of resources: liberty, for example (see *Sovereign Virtue*, chapter 3).

[32] Similarly, for reasons of simplicity I have chosen to measure along the axes of Figure 6 only one out of several possible sets of variables. A more sophisticated egalitarianism might prefer axes that measured equal opportunity for welfare and

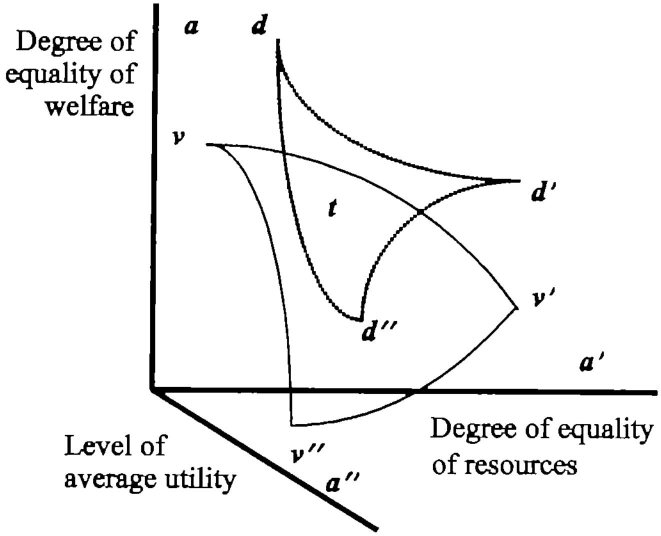

Degree of equality of welfare
a
d
v
t
d′
v′
d″
a′
Level of average utility
v″
a″
Degree of equality of resources

Figure 6: A three-dimensional indifference map.

So, Figure 6 simultaneously measures equality of welfare, equality of resources, and average utility.[35] I shall proceed to argue that some of the key features of egalitarian thinking are made salient by the touching at point *t* of the three-dimensional indifference plane $(dd'd'')$ and the three-dimensional availability plane $(vv'v'')$. But one thing is of immediate note. Even though Figure 6 represents the holding of three values that are not capable of simultaneous maximization, there is, nevertheless, a determinate point from which

---

equal opportunity for obtaining the resources one desires. Alternatively, if we are closer to what Parfit (*op. cit.*, p. 25) calls "nonrelational egalitarians" and subscribe to "the priority view," where equality is indirectly valued (because what really matters are benefits to the worst off), then the *Z*-axis might measure the extent to which the worst off are benefited.

[35] As Barry writes: "the principle of equal utility, if interpreted strictly, would require us to say that a situation of equally shared misery was morally preferable to a situation in which everyone was very happy but some were a little more so than others. It would seem more reasonable to say that we should be concerned with both the total amount of utility *and* its distribution, permitting a trade-off between the criteria of maximization and equalization"—*Theories of Justice* (Berkeley: California UP, 1989), p. 79. Clearly, as long as the population size is unaffected by the policy proposed, any policy that increases total utility will also increase average utility, and vice versa. Hence, excepting cases of variable population size, Barry's observation concerns average utility as much as it concerns total utility. I emphasize here average utility because of considerations to do with ascertaining the optimum population size—see "Moral Theory and Global Population."

other points may be viewed as undesirable, namely, when they are viewed from the standpoint of point *t*. In other words, value pluralism does not, as many presume, entail indeterminacy, nor does it entail an inability to generate substantive moral conclusions.

But is it really the case that all three axes are required if we are to shed light on egalitarian thinking? Dworkin's criticisms notwithstanding, it would seem that most egalitarians *do* value equality of welfare, for, as we have seen in the case of Sen's critique of Rawls, it may well be the possible resulting inequality in welfare which makes the Rawlsian distribution of primary social goods unappealing to many egalitarians.[34] Moreover, we have seen that even Dworkin acknowledges that equality of welfare holds considerable appeal for egalitarians. Thus, an axis measuring equality of welfare (or something similar) does, indeed, appear to be essential.

But, as Dworkin seems to have shown, it would also appear to be the case that most egalitarians value equality of resources. For it is the possibility of a highly unequal distribution of resources which, at least in part, makes the inclusion of certain deliberately cultivated, expensive tastes offensive to egalitarians. (This is represented by point *a* in Figure 6, which is a highly undesirable combination of distributions when viewed from the standpoint of the available combination represented by point *t*—point *t* falling both on the availability plane and on a far higher indifference plane than does point *a*.) Thus, an axis measuring equality of resources (or something similar) also seems to be necessary. But while equality of resources is often highly valued, it, too, can be unattractive—for example, when it leads to those who suffer greatly having their (less than incapacitating) pain ignored, and thus enjoying considerably less welfare than others (as represented by point *a′* in Figure 6, which is also a highly undesirable combination of distributions when viewed from the standpoint of point *t*).

Moreover, an axis measuring average utility also seems to be required, for when any attempt to attain equality of welfare significantly depresses average utility, it loses its attractiveness even to egalitarians—for example, consider the undesirability of redistributing to Dworkin's entirely paralyzed person the resources everyone else would need for a decent standard of living (back to point *a* in Figure 6). As this consideration has some purchase on egalitarians, it would appear that most of them ordinarily value a high level of average utility, as has already been argued. Nevertheless, it goes without saying that they are not solely concerned with a high level of average utility (represented by point $a''$ in Figure 6). One could imagine Sen's pleasure wizard having so much enjoyment that, on her own, she took any society of which she was a member to a high average utility level. A high level of average utility would then require an unequal distribution of both welfare and resources. But egalitarians would obviously condemn giving all the resources and all the welfare to the pleasure wizard.

---

[34] For a not wholly convincing attempt to deal with this objection, see Norman Daniels, "Equality of What: Welfare, Resources, or Capabilities?" *Philosophy and Phenomenological Research*, 1, Supplement (Fall 1990): 273-96.

In a word, their own respective arguments, and especially their reception, seem to establish beyond any reasonable doubt that most egalitarians tend to subscribe to *a plurality of values which*, most importantly, *are such that the maximizing of any one of them is not compossible with the maximizing of any other*. And this is what Figure 6 graphically portrays. Thus, it is hardly surprising that the search for the holy grail of a single metric by which to measure equality has proven so fruitless. There is no single currency of egalitarian justice.[35] Rather, as I suggested earlier, what we need is the equivalent of an egalitarian *bureau de change*. And that, in effect, is what a multidimensional indifference map constitutes.

Now, the three-dimensional map in Figure 6 makes explicit a fundamental problem for egalitarians: equality of resources and equality of welfare can pull in opposite directions[36] (as can equality of resources and the level of average utility, just as the level of average utility and equality of welfare can). But the three-dimensional map reveals something else besides. For while it is possible to trade equality of resources for equality of welfare, and vice versa, there are, nevertheless, points that fall well beneath an indifference plane (mapping the relations between the several values held by egalitarians) on which falls a feasible distribution (namely, point *t*). And clearly, were a substantial minority of people lacking in their basic capabilities, the situation would be represented by a point falling well beneath an available point on *any* reasonable and reflective egalitarian's preferred multidimensional indifference plane. Hence, an egalitarian society could reach general agreement in condemning situations where the basic capabilities of a substantial minority remained frustrated. Moreover, such agreement would be possible even if there were sufficient variations between individual multidimensional indifference planes to prevent agreement on what absolute equality consisted in. This means that agreement could be reached by egalitarians on social policy without their needing to order distributive principles lexically.

---

[35] Interestingly, in the second volume of his *Treatise on Social Justice*, Barry writes: "A number of attempts have been made to determine 'the currency of egalitarian justice'. Each of these is successful in pointing to examples that support the particular currency proposed and do not fit some rival currency. The common error lies in supposing that there is one answer waiting to be found...: justice is not to be defined in terms of the distribution of one fungible quantity"—*Justice as Impartiality* (New York: Oxford, 1995), p. 145.

[36] As Dworkin writes: "Equality is a popular but mysterious political ideal. People can become equal (or at least more equal) in one way with the consequence that they become unequal (or more unequal) in others. If people have equal income, for example, they will almost certainly differ in the amount of satisfaction they find in their lives, and vice versa"—"What Is Equality? Part 1: Equality of Welfare," p. 185.

But given the existence of a person so unfortunate that her basic capabilities could not be provided for adequately (as in the case of Dworkin's entirely paralyzed person—a person so unfortunate that endeavoring to relieve his misery significantly would lead to the average utility level plummeting), then the attempt to equalize everyone's welfare would, sadly, also fall beneath an available point on the preferred multidimensional indifference plane of any reasonable and reflective egalitarian. For the redistribution required to improve the unfortunate person's lot would result in an outcome containing (what would be viewed as) too little average utility, as well as too unequal a distribution of resources.[37]

Crucially, then, it is the distributions represented by points falling well beneath certain multidimensional indifference planes which egalitarians can be seen to oppose. Hence, it is perhaps more fruitful to conceive of "egalitarianism" not as the pursuit of equality, but as the opposition to certain *substantive inequalities*.[38] And it is not solely a greatly unequal distribution of resources, or welfare, or autonomy that triggers the concerns of the egalitarian. It is all of these inequalities—as well as low average utility. Moreover, and contra Sen, Dworkin, Arneson, and even Cohen,[39] any equal distribution along one dimension is bound to result in highly undesirable, substantive inequalities measured by the other metrics, which is why their search for a one-dimensional approach to equality seems ultimately doomed to failure.[40]

---

[37] Thus, given that there are likely to be limits to the amount of resources egalitarians would think it right to distribute to the worst off, then even subscribing to Parfit's "priority view" in preference to equality of welfare would still seem to require "nonrelational egalitarians" to engage in value trade-offs. Indeed, Parfit's "weighted beneficence" seems especially vulnerable to Dworkin's example of the entirely paralyzed person.

[38] Not surprisingly, then, I concur with much of what Baker argues (*op. cit.*, especially pp. 3-5).

[39] A significant difference between the position advocated here and Cohen's is that he sees responsibility (and, perhaps, ultimately autonomy) as *regulating* an egalitarian metric. Following Baker, I have argued that such an approach would result in unacceptable inequalities. Of course, an equal enjoyment of autonomy would inevitably lead to some inequalities in welfare and resources. But that means, on the approach advocated here, that trade-offs are therefore required (whereas Cohen's position views responsibility as, in effect, lexically prior to welfare and

## VII. CONCLUDING REMARKS

By way of conclusion, serious egalitarians are perhaps best viewed as most strongly motivated by an opposition to what they regard as *injustices*, rather than as being obsessed with the pursuit of some strict equality.[41] And what it is that is "egalitarian" about them, of course, is that the seeming injustices which most concern them involve substantive inequalities.

But which inequalities? On the whole, egalitarians oppose inequalities in (or unequal opportunities for) welfare, but not when too great an inequality in the distribution of resources or when too low an average utility would result from any attempt to correct them.[42] Equally, egalitarians, on the whole, oppose significant inequalities in the distribution of resources, but not when a consequence of their more equal distribution would be too great an inequality in (or too unequal an opportunity for) welfare, and not when it would result in

---

resource inequalities). Moreover, the approach presented here also implies a very different response to expensive tastes than Dworkin's While the valuing of autonomy will require some trade-offs with respect both to the equalizing of resources and to the equalizing of welfare such that at least some of the costs of one's choices will need to be borne by oneself, certain expensive tastes may well have to be acceded to (at least to some degree) because of the tremendous inequality in welfare that could result from simply ignoring them. But the drop in average utility and the inequality in resources which would result from their unrestricted satisfaction would together place considerable limits on acceding to extremely expensive tastes.

[40] It is not surprising, however, that they should all seek a single metric. Psychologists have accumulated considerable evidence indicating that human beings are somewhat limited in their ability to weigh a number of factors together while keeping the various trade-offs to the fore of their thinking. See Richard B. Brandt, *A Theory of the Good and the Right* (New York: Prometheus, 1998), p. 76.

[41] Thus, to employ Parfit's terminology once more, the majority of serious egalitarians are probably "deontic egalitarians" (*op. cit.*, pp. 3-4).

[42] For example, many egalitarians would surely agree with the following considered judgment: "It seems right...that a person who suffers from physical disabilities making it difficult and expensive for him to enjoy any of the normal pleasures of life should receive special assistance, and an adequate notion of equal treatment might well require such assistance. There may be controversies about the level of aid justified by such disabilities, and the principle of equal satisfaction may set this level too high, but the general tendency to recognize such claims does not seem wrong"—T. M. Scanlon, "Preference and Urgency," this JOURNAL, LXXII, 19 (November 6, 1975): 655-69, here p. 659. But it is unlikely that egalitarians would consider this general tendency to be right, while regarding equal satisfaction as too demanding, unless they valued equality of resources and/or a high level of average utility alongside equality of (or equal opportunity for) welfare.

too low an average utility.[43] Certain egalitarians would also rather see individuals equally autonomous, and thus equally living according to their own choices. But many would not want this to lead to great inequalities in either resource or welfare distribution.[44] And while there appears to be a common preference for a multidimensional equality, the difficulty (if not the impossibility) of achieving equality along all of the dimensions simultaneously brings to the fore what most seems to unite serious egalitarians, namely, their principled opposition to substantive inequalities, whether in the distribution of (or in the opportunities for) welfare, resources, or, perhaps, even autonomy.

ALAN CARTER

University of Colorado/Boulder

---

[43] The latter might result if monetary incentives were always necessary for increased material prosperity, which is moot—as is any assumption that such increases can always be equated with increased welfare.

[44] Compare: "A characteristic dispute about values in society, such as some issue of equality against freedom, is not one most typically enacted by a body of single-minded egalitarians confronting a body of equally single-minded libertarians, but is rather a conflict which one person equipped with a more generous range of human values, could find enacted in himself"—Bernard Williams, "Conflicts of Values," in *Moral Luck* (New York: Cambridge, 1981), pp. 71-82, here p. 73.

## BOOK REVIEWS

*Culture and Equality.* BRIAN BARRY. Cambridge: Harvard University Press, 2001. xi + 399 p. Cloth $36.95.

One of the most persistent criticisms of liberalism is that the priority it assigns to freedom and individual rights is not simply disruptive of conventional social norms but also undermines the values of community. The communitarianism which arose in the 1980s is a recent example. Some communitarians are more liberal than others (for example, Michael Walzer and Charles Taylor more so than Alasdair MacIntyre[1]). Still, communitarians are united around the conviction that basic freedoms and other requirements of (liberal) justice are secondary to achieving the values of community. Most liberals can accept that individuals realize a large part of their good through participation in social groups (families and friendships, and larger associations too); these and other values of community are worth pursuing for their own sake. Liberals, however, reject the communitarian view that comprehensive communal purposes are to be *politically* enforced, taking priority over equal basic liberties and opportunities and the freedom to define one's own good.

Multiculturalism is the heir to this nonliberal communitarian doctrine. (It is no accident that many communitarians also profess multiculturalism.) Like communitarians, multiculturalists insist that a person's good is primarily defined by membership and active participation in a (dominant) community. But whereas communitarianism outlines the bases of social unity in terms of everyone's adhering to shared norms that achieve communal ends, multiculturalism takes cognizance of the fact that often there are a multiplicity of cultures co-existing within the same society and under one government. It then provides communitarianism with a nonideal theory which says how societies and their governments should deal with the real world of "difference."[2] Multiculturalists contend that, because achieving

---

[1] Walzer, *Spheres of Justice: A Defense of Pluralism and Equality* (New York: Basic, 1983); Taylor, "Cross-Purposes: The Liberal Communitarian Debate," in his *Philosophical Arguments* (Cambridge: Harvard, 1995), pp. 181-203; MacIntyre, *After Virtue* (Notre Dame: University Press, 1984).

[2] Another comparison emphasizes the contrast between views: while communitarians tend to deny social diversity for the sake of shared norms that achieve communal ends, multiculturalists tend to celebrate difference and deny the possibility of universal principles or shared ends. See Jon Mandle, *What's Left of Liberalism* (Lanham, MD: Lexington, 2000), pp. 6-8, 10.

0022-362X/02/9911/600–06

one's cultural "identity" is so central to a person's good, each distinct cultural group in a multicultural society should recognize and respect the cultural practices of others and not impose its norms, particularly its liberal norms, on them. For the liberal emphasis on individuals' equal freedom to find their own good makes achieving one's cultural identity difficult, if not practically impossible, and undermines the distinctness of cultural groups. Multiculturalism prescribes then a policy not just of toleration but of accommodation of disparate cultural groups, many of which do not endorse liberal social or even political norms.

Brian Barry's *Culture and Equality* is a liberal response to multiculturalism and its criticisms of liberalism. The book is a sustained attack on multiculturalism's main theses and proponents, from the perspective of the kind of egalitarian liberalism expressed in John Rawls's *A Theory of Justice*[3] (7f., 16). Barry has been one of Rawls's more informed and probing liberal critics. But for all his differences with Rawls, Barry still sees justice as fairness as the major statement of "the classical ideal of liberal citizenship" and the egalitarian "demands of social and economic citizenship" that define egalitarian liberalism (7). Barry sees multiculturalists as denying the equality of basic liberties and fair opportunities that define equal liberal citizenship; moreover, they even undermine the economic claims of the poor by trying to shift political focus away from questions of distributive justice to a "politics of recognition" of different cultural groups (325).

Barry's attitude toward multiculturalism is suggested early on: "I have found that there is something approaching a consensus among those who do not write about it that the literature of multiculturalism is not worth wasting powder and shot on" (6). What makes critical engagement with multiculturalism worthwhile for Barry is that it receives such widespread sympathy, if not allegiance, in academia and American intellectual life, and those who write on multiculturalism are almost uniformly sympathetic to it. While nonphilosophers (such as Robert Hughes and Todd Gitlin[4]) have responded to multiculturalism on behalf of liberalism, Barry sees a sustained critical treatment from within political philosophy as long overdue.

*Culture and Equality* is divided into three parts, according to topics. In Part I, Barry concentrates on the idea of equal treatment. He takes on the multiculturalist view that equal treatment requires treating people according to their different culturally derived beliefs and practices. He basically argues that this misconstrues the liberal ideal of equal treatment, which requires that people be treated according to the same rules and requirements. Part II focuses on the claims of groups and particularly the multiculturalist idea of "group rights." Barry addresses the argument that liberal principles tend to undermine or destroy the independence of minority cultures and that these groups should have special group rights to protect their cultural practices. He focuses especially on the claims of illiberal religions (for example, the Amish) for special protection, and also addresses religious practices of sex discrimination (chapter 5). Then he turns to religious and other groups' claims regarding the rearing and education of children (chapter 6). In Part III, Barry addresses philosophical arguments for multiculturalism that proceed from the idea that moral universalism is false (chapter 7). The book ends with a discussion of the adverse practical consequences of enacting multiculturalist programs (chapter 8). Here, Barry argues mainly that these policies do not benefit the people they are intended to help, and prevent the enactment of liberal social programs that really would benefit the disadvantaged members of minority groups.

---

[3] Cambridge: Harvard, 1971.

[4] Hughes, *Culture of Complaint: The Fraying of America* (New York: Oxford, 1993); Gitlin, *The Twilight of Common Dreams: Why America is Wracked by Culture Wars* (New York: Holt, 1995).

Liberalism for Barry is a universal doctrine that applies to all persons in all societies as a matter of right and justice. It requires equal basic liberties, a strong view of equal opportunities, and guaranteed economic resources for all persons. Societies that are not sufficiently developed to provide all these rights and goods still have a duty to work toward institutions that eventually will. Barry will have nothing to do with Walzer's and other multiculturalists' position that the rights and liberties people ought to have depend on the "shared understandings" or practices of their cultures (136). This is cultural relativism and it is ultimately incoherent (chapter 7). Moreover, it makes justice dependent on the values and views of dominant elites, and discriminates against minorities who do not share the understandings of a majority (196).

Barry sees multiculturalism as regressive. It is "anti-egalitarian," if not in intention, then certainly in effect (12). The privileges it provides to special interests are "conducive to a politics of 'divide and rule' that can only benefit those who benefit most from the status quo" (11). He finds it especially ironic that the multiculturalist left would seek to revive the romantic doctrine that each cultural group has an identity uniquely suited to it which ought to be preserved, cultivated, and if necessary even resuscitated. Since cultural identity is not chosen but is largely based on descent, the multiculturalist left's embrace of "romantic nationalism" flirts with the worst twentieth-century right-wing ideologies (260-61).

Barry sees equal treatment as an integral feature of liberalism. Equal treatment does not imply equal impact, he says, but governing everyone according to the same legal rules. Almost any law will affect people differently, and by itself there is nothing inherently unfair about this (34). He rejects then the "rule-and-exemption" approach to religious and other minorities advocated by multiculturalists, which exempts minority practices from general legal requirements. So he opposes the exceptions made in Britain's animal-slaughter laws which allow Jews and Muslims to use traditional methods of ritual slaughter, as well as Britain's exemption for Sikhs from weapons and motorcycle helmet laws (41-46). He further argues that the Supreme Court, in *Oregon v. Smith*, correctly denied members of the Native-American Church the right to use peyote sacramentally in exception to antidrug laws, for it would violate liberal equal treatment constitutionally to require an exemption for religious use of illegal drugs (170ff., 183f.).

Here Barry exaggerates the degree to which the liberal principles of justice he relies on always require equal treatment according to the same legal rules. Sometimes there are legitimate liberal objections to equal treatment according to a legal rule. Equal treatment under one rule may involve, not just unequal impact, but unequal treatment under another rule. Then the important question for liberals is not whether equal treatment damages someone's cultural identity, but whether some important right or other requirement of justice is violated. Given the priority liberals assign to equal liberty of conscience, the Supreme Court should have made an exception in *Smith* to drug laws for sacramental practices that themselves do not violate anyone's basic rights or other important requirements of justice. No doubt the Court (and particularly Justice Antonin Scalia, who wrote *Smith*) would not have enforced a legal prohibition on alcohol against the use of wine in Catholic liturgy. There is no difference with the sacramental use of peyote (assuming it is a central part of the Native American Church's liturgy). In *Smith*, equal treatment under drug laws resulted in unjust (not simply unequal, though that too) treatment under the First Amendment.

The priority given to such basic equal rights as liberty of conscience is the main way that liberalism protects the integrity of minority groups and cultural practices. Barry argues (chapter 4) that freedom of association especially provides for a liberal culture of tolerance which allows for the diversity and flourishing of many different cultural traditions. Freedom of association involves the authority of groups to treat their members in illiberal ways, as unequals (for example, religious and other traditional restrictions on women's roles), and by limiting individual freedom as a condition of membership (for example, religious dietary and sexual conduct restrictions). "It is no part of liberalism...to insist that every group must conform to liberal principles in its internal structure" (147). As long as restrictions on conduct are voluntarily assumed by members, there is no violation of liberal political norms. But (many) multiculturalists reject liberal diversity and argue instead for "deep diversity," which involves imposing coercive political restrictions on the liberties of members of minority cultures, to prevent them from deviating from cultural norms (128). Barry perceptively explains how such proposed restrictions imply a rejection of liberal freedom of association. Essential to this liberty is a person's right to refuse group demands and exit associations (149f.). This is one way liberals differ significantly not just from many multiculturalists, but also from libertarians. No court in a liberal society would enforce an agreement whereby a person as a condition of membership voluntarily alienated her right to change her religious or other group affiliation, or gave up her right to divorce. The rights that constitute freedom of association are (like other basic liberties) inalienable.

Barry's most trenchant criticisms are directed against liberals who seek to accommodate multicultural aspirations. Will Kymlicka[5] especially is criticized, since "he presents himself as [a liberal]" (137), but in fact he is not since he would compromise liberalism's universalistic and egalitarian core in the name of the "romantic nationalism" that he advocates (along with Taylor and multiculturalists generally). Barry bases these criticisms on Kymlicka's (and Walzer's) willingness to grant national minorities (such as Quebec) rights of self-government within a liberal constitution that allow them to make exceptions "to measures imposed by a liberal state to prevent violations of liberty and equality" (138). Barry cites (as an example) Kymlicka's willingness to allow Pueblo tribal councils the authority to limit freedom of conscience and impose sexually discriminatory political membership rules. Against Kymlicka's contention that his "asymmetric federalism" does not involve any inequality for Canadians outside Quebec, Barry contends that Kymlicka ignores the obvious inequality that allows Quebec representatives to vote on laws that apply to the rest of Canada (ROC) but not to Quebec (311). In fairness, Kymlicka does object to the Pueblo rejection of liberal liberties previously mentioned, but opposes the United States government's "imposing liberalism" by coercing the Pueblo council (*ibid.*, p. 165). Since Kymlicka sees a Pueblo violation of rights here, it perhaps presumes too much to say that "Kymlicka clearly buys into the idea that human rights are a form of 'cultural imperialism'" (138). This is especially so if not all equal liberal liberties (for example, equal political rights to vote and hold office) are also human rights.

---

[5] *Multicultural Citizenship* (New York: Oxford, 1995).

A philosophical notion of mutual respect for persons (and citizens) informs liberal thinking and is one basis for the idea of human (and liberal constitutional) rights. The liberal idea of respect is different from the idea of "recognition" that attracts multiculturalists (for example, Taylor, Iris Young, Nancy Fraser, and James Tully, among others Barry discusses). The "politics of recognition" does not deny the universalist idea of respect for persons as such; rather, it insists that a condition of equal respect is that the diverse cultural practices and values affirmed by different persons receive recognition of their equal worth. Barry finds this position incoherent. "Unless discriminations are made, ascribing value to something ceases to have any point" (269). Liberal requirements of equal respect, equal treatment, and equal rights are *political* duties of justice morally owed to persons, and do not depend on recognition of the equal worth of their lifestyle. So in response to Andrew Sullivan's[6] and others' arguments, Barry says: "We should totally reject the notion that the only way in which the case for equal rights for homosexuals can be made is to establish first the equal worthiness of homosexual and hetereosexual lifestyles" (279). It is not only bad philosophy in Barry's view, but also a self-defeating political strategy (276-77).

It may be that Barry can be criticized for not being sufficiently attuned to some of multiculturalists' main concerns. They are not just worried that the distinctness of cultures and cultural practices will be lost in a liberal society, but that they are being melded into ways of life typical of middle America as it responds to the growing influence of global capitalism. One does not have to be a romantic nationalist to regret the effects of popular culture (including commercial television) in homogenizing life and undermining culturally and regionally distinct ways of life. The ever present depiction of brutality and carnality by the entertainment industry is not a problem liberalism can easily address politically. Liberalism leaves it up mainly to families to exercise control in these matters. But multiculturalists know this is something most parents, even when they are able, fail to do, however much parents may regret what their children are exposed to.

---

[6] "What's So Bad about Hate?" *New York Times Magazine*, xxvi (September 1999): 50-57.

Barry is right to say that liberalism does not seek to meld all cultures into a distinct pattern. On the contrary, liberalism (unlike other political views) respects cultural differences (so long as they respect the liberal political and civil rights of their members) by allowing freedom of association and other liberties needed for a distinct culture to survive in a diverse society. What liberalism refuses to do is to ensure a culture's survival by enforcing politically the practices of any particular cultural group. Instead, it politically permits individuals to revise their "cultural identities." The consequence of this is the thriving (and dying) of many different cultural groups, accompanied by a largely traditionless and commercialized "civil society" to which individuals turn, often to escape the confines of their particular cultures. It is perhaps the homogenizing effects of liberal civil society, seen by multiculturalists as largely individualistic and commercially infused, which they object to most. While its enormous commercial influences might be regulated, this individualistic mass culture is perhaps an ineradicable part of liberalism. G. W. F. Hegel proposed the unifying forces of "the State" as a source of community to temper the individualistic and commercialized bias of liberal civil society. Liberals (fortunately) do not have that option. Government's role is not to enforce a community of (nonpolitical) values, but to establish justice and promote the common good of free and equal citizens. This is the common liberal political culture that provides the basis for social unity among disparate subcultures and groups. Still a problem (if that is what it is) remains, and liberalism's response to it (thus far) is unsatisfactory for many people. Barry thoroughly and effectively criticizes all the illicit (and even a few licit) ways that multiculturalists seek to address the problems they see in liberal political and social culture. Hopefully, he and other liberals[7] might now devote greater attention to liberal means that counteract the disintegrating effects of commercial and other institutions in liberal civil society.

SAMUEL FREEMAN

University of Pennsylvania

---

[7] One such effort is Joseph Raz's "Multiculturalism: A Liberal Perspective," in his *Ethics in the Public Domain* (New York: Oxford, 1994), pp. 170-91.

Asking **Why**

A leading philosopher asks: Did we have a responsibility for what took place on September 11? Did we respond to it as we should have? What are we to do now?

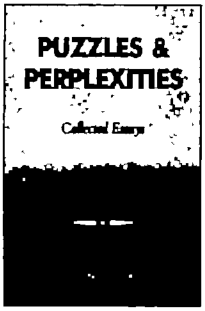
**"In the flood of literature on terrorism, this essay stands out as unusual, and unusually enlightening and provocative."** —Noam Chomsky

**"[A] rare delight. This uncompromising and courageous philosopher continues the dissenting tradition of Bertrand Russell and Jean-Paul Sartre, more needed now than ever before."** —Tariq Ali

**AFTER THE TERROR** BY TED HONDERICH • $22.50 cloth

**EDINBURGH UNIVERSITY PRESS**
Distributed by Columbia University Press • columbia.edu/cu/cup • 800-944-8648

---



Puzzles & Perplexities
*Collected Essays*
By Steven M. Cahn

"*Puzzles & Perplexities* is a thought-provoking collection of Steven Cahn's far-reaching writings, ranging over such diverse topics as free will, the Book of Job, affirmative action, and divestiture. Each essay raises a problem, whether it be in metaphysics, ethics, philosophy of religion, or social policy that will not only engage and stimulate students, but also challenge scholars to examine their own views as well."
—**Robert Simon, Hamilton College**

"Professor Cahn's essays make the mystery of philosopy accessible as he considers many of the important intellectual issues of our time."
—**Norman Bowie, University of Minnesota**

"In *Puzzles & Perplexities*, Steven Cahn addresses questions about faith and rationality, social and aesthetic value, and the theory and practice of education. He writes lucidly and passionately about problems that are both timely and timeless, about issues that should engage all thinking persons." —**Karen Hanson, Indiana University**

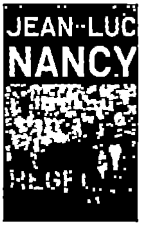
October 2002 • 152 pages • Paper 0-7425-1423-4 $26.95 • Cloth 0-7425-1422-6 $75.00
To order call: 1-800-462-6420 or visit www.RowmanLittlefield.com

Rowman & Littlefield Publishers, Inc.

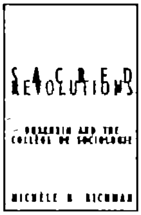
Thought-Provoking

VILÉM FLUSSER
Writings
Translated by Erik Eisel
Edited by Andreas Strohl
The first English-language anthology from one of Europe's most influential, original, and provocative theorists
$29.95 cloth/jacket • 256 pages
Electronic Mediations Series, volume 6

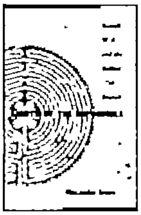
EDWARD S. CASEY
Representing Place
*Landscape Painting and Maps*
Edward S. Casey combines history with philosophy, and cartography with art, to create a new understanding of how representation requires and thrives on space.
$29.95 paper • $82.95 cloth • 392 pages • 16 color photos, 38 halftones, 14 illustrations



JEAN-LUC NANCY
Hegel
*The Restlessness of the Negative*
Translated by Jason Smith and Steven Miller
A major philosopher gives us a new view of Hegel for our time.
"Nancy makes Hegel accessible, but he also makes him new, evoking a sense of the affirmative and open character of his thinking that is rarely acknowledged within contemporary discourse." —Judith Butler
$17.95 paper • $49.95 cloth • 168 pages



MICHÈLE H. RICHMAN
Sacred Revolutions
*Durkheim and the Collège de Sociologie*
Explores the role of the sacred in the response of French intellectuals to the rise of fascism during the 1930s.
$19.95 paper • $54.95 cloth • 256 pages
Contradictions Series, volume 14



ALEXANDER IRWIN
Saints of the Impossible
*Bataille, Weil, and the Politics of the Sacred*
A surprising exploration of the relationship between these two twentieth-century thinkers.
$19.95 paper • $54.95 cloth • 256 pages

GEOFFREY A. HALE
Kierkegaard and the Ends of Language
A timely and original intervention in our understanding of this major philosopher through the lens of his influence on Franz Kafka, Theodor Adorno, and Rainer Maria Rilke.
$19.95 paper • $54.95 cloth • 264 pages

University of Minnesota Press
www.upress.umn.edu
773-568-1550
MINNESOTA

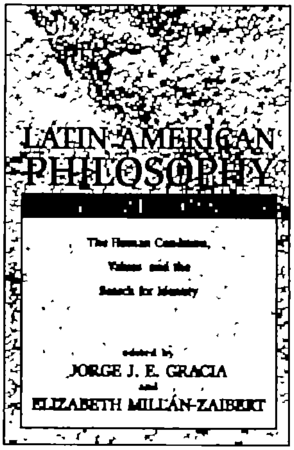
*New Philosophy Titles from . . .*

Prometheus Books

**LATIN AMERICAN PHILOSOPHY FOR THE 21ST CENTURY**
The Human Condition, Values, and the Search for Identity
**Edited by Jorge J. E. Gracia and Elizabeth Millán-Zaibert**
250 pp • ISBN 1-57392-978-6 • PB $28

**THE PHILOSOPHICAL DICTIONARY**
Enlarged Edition
**Mario Bunge**
330 pp • ISBN 1-59102-037-9 • PB $24

**HAPPINESS IN A GODLESS WORLD**
**George H. Smith**
250 pp • ISBN 1-59102-038-7 • PB $19


LATIN AMERICAN PHILOSOPHY
The Human Condition, Values and the Search for Identity
edited by
JORGE J. E. GRACIA
and
ELIZABETH MILLÁN-ZAIBERT

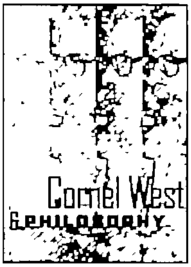
*. . . and*

Humanity Books

**THE HEIDEGGER-JASPERS CORRESPONDENCE**
(1920–1963)
**Edited by Walter Biemel and Hans Saner**
*Translated by Gary E. Aylesworth*
280 pp • ISBN 1-59102-060-3 • HC $59

**HEIDEGGER, DECISIONISM, AND QUIETISM**
Mark Basil Tanzer
119 pp • ISBN 1-59102-061-1 • HC $50

For exam copies contact editorial@prometheusbooks.com
Call toll free (800) 421-0351 • Fax (716) 691-0137 • www.prometheusbooks.com



ROUTLEDGE

NEW IN PHILOSOPHY

**THE AFFIRMATIVE ACTION DEBATE**
Second Edition
*Edited by Steven M. Cahn*
This newly updated, highly readable collection provides an essential guide to the full range of arguments surrounding contemporary discussion of affirmative action
**248 pp $22.95**

**CORNEL WEST AND PHILOSOPHY**
*Clarence Shole Johnson*
In this thought-provoking book, Johnson provides a unique and indispensable guide to West's diverse philosophical writings.
**240 pp $24.95**

**LATER DERRIDA**
Reading the Recent Work
*Herman Rapaport*
*"...proposes insightful interpretations of Derrida's latest work in the domain of politics and ethics. In the process, it challenges some of the most entrenched ideas about deconstruction."*
—Peggy Kamuf
**168 pp $19.95**

**LYOTARD**
Philosophy, Politics and the Sublime
*Edited by Hugh J. Silverman*
*Lyotard* is a thorough reassessment of Lyotard's extraordinary legacy and contribution to contemporary cultural, political, ethical, and aesthetic theory, and is an indispensable guide to key issues in his philosophy
**304 pp $24 95**

**PHILOSOPHY OF SCIENCE AND RACE**
*Naomi Zack*
In this concisely argued book, Zack explores the scientific and philosophical problems in applying a biological conception of race to human beings
**160 pp $19.95**

**THE SELECTED LETTERS OF BERTRAND RUSSELL**
The Private Years, 1884-1914
The Public Years, 1914-1970
*Edited by Nicholas Griffin*
*"Anybody who reads this splendid and revealing selection of letters will soon feel that he knows Russell. Nicholas Griffin completes a two-volume epistolary biography unrivaled for its scholarship in the over-populated world of Russell studies"*
—*The Economist*
**Vol I: 568 pp $18.95**
**Vol II: 688 pp $18.95**

1.800.634.7064
www.routledge-ny.com

Routledge
Taylor & Francis Group

# OUTSTANDING SCHOLARSHIP FROM CAMBRIDGE

**A Defense of Abortion**
*David Boonin*
Cambridge Studies in Philosophy and Public Policy
0-521-81701-3 Hb $65.00
0-521-52035-5 Pb $23.00

**Women Philosophers of the Seventeenth Century**
*Jacqueline Broad*
0-521-81295-X Hb $60.00*

**The Cambridge Companion to Rawls**
*Samuel Freeman, Editor*
Cambridge Companions to Philosophy
0-521-65167-0 Hb $65.00
0-521-65706-7 Pb $24.00

**Immanuel Kant**
A Biography
*Manfred Kuehn*
0-521-52406-7 Pb $25.00

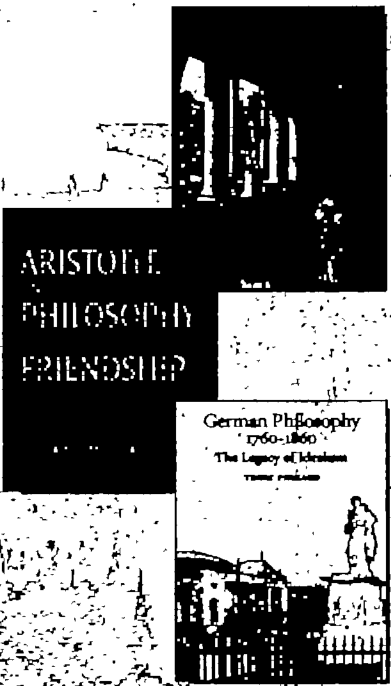
**Free Logic**
Selected Essays
*Karel Lambert*
0-521-81816-8 Hb $55.00

**Thomas Kuhn**
*Thomas Nickles, Editor*
Contemporary Philosophy in Focus
0-521-79206-1 Hb $58.00
0-521-79648-2 Pb $21.00

**Aristotle and the Philosophy of Friendship**
*Lorraine Pangle*
0-521-81745-5 Hb $65.00

**German Philosophy 1760-1860**
The Legacy of Idealism
*Terry Pinkard*
0-521-66326-1 Hb $65.00
0-521-66381-4 Pb $23.00

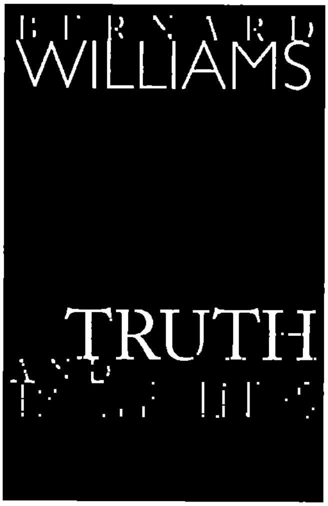
**The Ethics of Deference**
Learning from Law's Morals
*Philip Soper*
Cambridge Studies in Philosophy and Law
0-521-81047-7 Hb $60.00
0-521-00872-7 Pb $23.00



*At better bookstores*

CAMBRIDGE
UNIVERSITY PRESS

www.cambridge.org



# New from Princeton

**Truth and Truthfulness**
An Essay in Genealogy
**Bernard Williams**

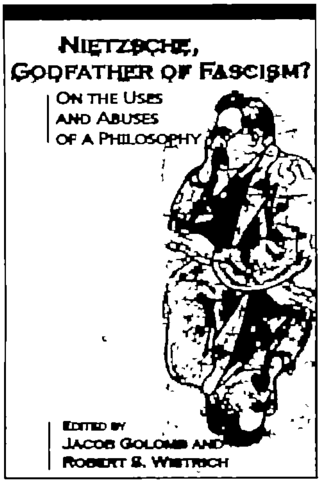
What does it mean to be truthful? What role does truth play in our lives? What do we lose if we reject truthfulness? No philosopher is better suited to answer these questions than Bernard Williams. Writing with his characteristic combination of passion and elegant simplicity, he explores the value of truth and finds it to be both less and more than we might imagine.

*Truth and Truthfulness* presents a powerful challenge to the fashionable belief that truth has no value, but equally to the traditional faith that its value guarantees itself. Bernard Williams shows us that when we lose a sense of the value of truth, we lose a lot both politically and personally, and may well lose everything.

**"In this exceptionally brilliant book ... Bernard Williams daringly asks—and still more daringly answers—one of the central questions of philosophy: what is the point of telling the truth? Lucid, penetrating, and profound, Williams' reflections are vitally important not for philosophers alone but for anyone interested in human thought and creativity."**
—Stephen Greenblatt,
Harvard University

Cloth $27.95 ISBN 0-691-10276-7

New from Princeton


Nietzsche, Godfather of Fascism?
On the Uses and Abuses of a Philosophy
Edited by Jacob Golomb and Robert S. Wistrich

**Nietzsche, Godfather of Fascism?**
On the Uses and Abuses of a Philosophy
**Edited by Jacob Golomb and Robert S. Wistrich**

The question that lies at the heart of this collection is how Nietzsche came to acquire the deadly "honor" of being considered the philosopher of the Third Reich and whether such claims had any justification. Contributions by a variety of distinguished philosophers and historians explore in depth Nietzsche's attitudes toward Jews, Judaism, Christianity, anti-Semitism, and National Socialism.

Paper $16.95 ISBN 0-691-00710-1
Cloth $65.00 ISBN 0-691-00709-8

**Praise and Blame**
Moral Realism and Its Applications
**Daniel N. Robinson**

In this lively book, Daniel Robinson takes on the prodigious task of setting forth the contours of praise and blame. He does so by mounting an important and provocative new defense of a radical theory of moral realism and offering a critical appraisal of prevailing alternatives such as determinism and behaviorism and of their conceptual shortcomings. Ultimately, the theory that Robinson develops preserves moral properties while acknowledging the conditions that undermine the powers of human will.

"*Praise and Blame* is a terrific book.... People who are looking for an original take on perennial philosophical questions won't be disappointed."
—Robert P. George, Princeton University

*New Forum Books*
Cloth $29.95 ISBN 0-691-057-9

**Princeton University Press**

*New from*
**Oxford**

**Human Welfare and Moral Worth**
Kantian Perspectives
THOMAS E. HILL
Thomas Hill, a leading figure in the recent development of Kantian moral philosophy, presents a set of essays that will be indispensable to moral, political, and social philosophers in exploring the implications of basic Kantian ideas for practical issues.
paper $19.95 cloth $55.00

*New in paperback!*
**Essays on Indian Philosophy**
Traditional and Modern
J. N. MOHANTY
Edited by PURUSHOTTAMA BILIMORIA
"This outstanding anthology provides an excellent introduction to Indian culture, spirituality, and history. The book is an invaluable teaching resource for classes in Asian studies and serves as a standard reference for Indian culture. Very highly recommended."—*Religious Humanism.*
paper $19.95

**Abortion and Social Responsibility**
Depolarizing the Debate
LAURIE SHRAGE
Schrage argues that *Roe v Wade's* regulatory scheme of a six-month time span for abortion on demand polarized the public and obscured alternatives with potentially broader support. She explores the origins of that scheme, then defends an alternate one—with a time span shorter than 6 months for non-therapeutic abortions—that could win broad support needed to make legal abortion services available to all women.
paper $19.95 cloth $60.00

**Autonomy, Gender, Politics**
MARILYN FRIEDMAN
Marilyn Friedman defends the importance of autonomy in romantic love, considers how social institutions should respond to women who choose to remain in abusive relationships, and argues that liberal societies should tolerate minority cultural practices that violate women's rights so long as the women in question have chosen autonomously to live according to those practices.
paper $19.95 cloth $60.00

**In Gods We Trust**
The Evolutionary Landscape of Religion
SCOTT ATRAN
A cognitive anthropologist and psychologist, Scott Atran argues that religion is a by-product of human evolution just as the cognitive intervention, cultural selection, and historical survival of religion is an accommodation of certain existential and moral elements that have evolved in the human condition.
$35.00

**The Compositionality Papers**
Edited by JERRY A. FODOR and ERNIE LEPORE
Lepore and Fodor present a collection of original and controversial essays that will prove essential reading for anyone interested in issues relating to compositionality in language and mind.
paper $19.95 cloth $55.00

Prices are subject to change and apply only in the US. To order, or for more information, please call 1-800-451-7556.
Visit our website at www.oup-usa.org

OXFORD

# THE JOURNAL OF PHILOSOPHY

02\4\06

VOLUME XCIX, NUMBER 12
DECEMBER 2002

page
607 *Self-locating Belief in Big Worlds: Cosmology's Missing Link to Observation* Nick Bostrom
624 *Belief and Translation* Marco Santambrogio

BOOK REVIEWS
648 Daniel Garber: *Descartes Embodied: Reading Cartesian Philosophy through Cartesian Science* John Cottingham

651 NOTES AND NEWS
652 INDEX TO VOLUME XCIX

*Published by* **The Journal of Philosophy, Inc.**

# THE JOURNAL OF PHILOSOPHY

FOUNDED BY FREDERICK J. E. WOODBRIDGE AND WENDELL T. BUSH

Purpose: To publish philosophical articles of current interest and encourage the interchange of ideas, especially the exploration of the borderline between philosophy and other disciplines.

*Editors*: Bernard Berofsky, Akeel Bilgrami, Arthur C. Danto, Kent Greenawalt, Patricia Kitcher, Philip Kitcher, Isaac Levi, Mary Mothersill, Philip Pettit, Carol Rovane, Achille C. Varzi. *Editor Emeritus*: Sidney Morgenbesser. *Consulting Editors*: David Albert, John Collins, James T. Higginbotham, Charles D. Parsons, Wilfried Sieg. *Managing Editor*: John Smylie.

THE JOURNAL OF PHILOSOPHY is owned and published by the Journal of Philosophy, Inc. *President*, Arthur C. Danto; *Vice President*, Akeel Bilgrami; *Secretary*, Daniel Shapiro; *Treasurer*, Barbara Gimbel; *Other Trustees*: Lee Bollinger, Leigh S. Cauman, Kent Greenawalt, Michael J. Mooney, Lynn Nesbit.

All communications to the Editors and Trustees and all manuscripts may be sent to John Smylie, Managing Editor, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, New York 10027. FAX: (212) 932-3721.

You may also visit our website at: www.journalofphilosophy.org

# THE JOURNAL OF PHILOSOPHY

2003

SUBSCRIPTIONS (12 issues)

| | |
|---|---|
| Individuals | $35.00 |
| Libraries and Institutions | $75.00 |
| Students, retired/unemployed philosophers | $20.00 |
| Postage outside the U.S. | $15.00 |

Payments only in U.S. currency on a U.S. bank. All back volumes and separate issues available back to 1904. Please inquire for price lists, shipping charges, and discounts on back orders. Please inquire for advertising rates; ad space is limited, so ad reservations are required.

Published monthly as of January 1977; typeset and printed by Cadmus Journal Services, Lancaster and Ephrata, PA.

All communication about subscriptions and advertisements may be sent to Pamela Ward, Business Manager, Mail Code 4972, 1150 Amsterdam Avenue, Columbia University, New York, NY, 10027. (212) 866-1742

The JOURNAL allows copies of its articles to be made for personal or classroom use, *if* the copier abides by the JOURNAL's terms for all copying beyond that permitted by Sections 107 or 108 of the U.S. Copyright Law. This consent does not extend to any other kinds of copying. More information on our terms may be obtained by writing to us.

POSTMASTER: Periodical postage paid at New York, NY, and other mailing offices.
POSTMASTER: Send address changes to the *Journal of Philosophy* at MC 4972, Columbia University, 1150 Amsterdam Avenue, New York, NY 10027.

© Copyright 2002 by the Journal of Philosophy, Inc.

ISSN 0022-362X

THE JOURNAL OF PHILOSOPHY
VOLUME XCIX, NO. 12, DECEMBER 2002

62\9\03

# SELF-LOCATING BELIEF IN BIG WORLDS: COSMOLOGY'S MISSING LINK TO OBSERVATION*

Space is big. It is very, *very* big. On the currently most favored cosmological theories, we are living in an infinite world, a world that contains an infinite number of planets, stars, galaxies, and black holes. This is an implication of most "multiverse theories," according to which our universe is just one in a vast ensemble of physically real universes. But it is also a consequence of the standard Big Bang cosmology, if combined with the assumption that our universe is open or flat, as recent evidence suggests it is. An open or flat universe—assuming the simplest topology[1]—is spatially infinite at any time and contains infinitely many planets, and the like.[2]

---

* I am grateful for valuable comments from Craig Callender, Milan Cirkovic, Adam Elga, Colin Howson, John Leslie, Peter Milne, Don Page, Elliott Sober, Alex Vilenkin, and Roger White.

[1] That is, that space is simply connected. There has been a recent burst of interest in the possibility that our universe might be multiply connected, in which case it could be both finite and hyperbolic. A multiply connected space could lead to a telltale pattern consisting of a superposition of multiple images of the night sky seen at varying distances from Earth (roughly, one image for each lap around the universe that the light has traveled). Such a pattern has not been found, although the search continues. For an introduction to multiply connected topologies in cosmology, see M. Lachièze-Rey and J.-P. Luminet, "Cosmic Topology," *Physics Reports*, CCLIV, 3 (1995): 135-214.

[2] A widespread misconception is that the open universe in the standard Big Bang model becomes spatially infinite only in the temporal limit. The *observable* universe is finite, but only a small part of the whole is observable (by us). One fallacious intuition that might be responsible for this misconception is that the universe came into existence at some spatial point in the Big Bang. A better way of picturing things is to imagine space as an infinite rubber sheet, and gravitationally bound groupings (such as stars and galaxies) as buttons glued on to it. As we move forward in time, the sheet is stretched in all directions so that the separation between the buttons

0022-362X/02/9912/607–23

Philosophical investigations relating to the vastness of the cosmos have focused on the fine-tuning of our universe. 'Fine-tuning' refers to the alleged fact that the laws of physics are such that, if any of several physical constants had been even slightly different, then life would not have existed. A philosophical cottage industry has arisen from the controversies surrounding issues, such as whether fine-tuning is in some sense "improbable," whether it should be regarded as surprising,[3] whether it calls out for explanation (and if so whether a multiverse theory could explain it),[4] whether it suggests ways in which current physics is incomplete,[5] or whether it is evidence for the hypothesis that our universe was designed.[6]

Here, I wish instead to address a more fundamental problem: How can vast-world cosmologies have *any* observational consequences *at all*? I shall show that these cosmologies imply, or give a very high probability to, the proposition that every possible observation is in fact made. This creates a challenge: If a theory is such that for any possible human observation that we specify, the theory says that that observation will be made, then how do we test the theory? What could possibly count as negative evidence? And if all theories that share this feature are equally good at predicting the data we shall get, then how can empirical evidence distinguish among them?

I call this a "challenge," because current cosmological theories clearly do have connections to observation. Cosmologists are constantly modifying and refining theories in light of empirical findings, and they are presumably not irrational in doing so. But it is a philosophical problem to account for how this is possible.

One lesson that will emerge is that we must be careful about how we construe the evidence. We know not only that such-and-such

---

increases. Going backward in time, we imagine the buttons coming closer together until, at "time zero," the density of the (still spatially infinite) universe becomes infinite everywhere. See, for example, J. L. Martin, *General Relativity* (New York: Prentice Hall, 1995).

[3] For example, John Earman, "The SAP Also Rises: A Critical Examination of the Anthropic Principle," *Philosophical Quarterly*, XXIV, 4 (1987): 307-17; John Leslie, *Universes* (New York: Routledge, 1989).

[4] For example, Quentin Smith, "Anthropic Explanations in Cosmology," *Australasian Journal of Philosophy*, LXXII, 3 (1994): 371-82; Ian Hacking, "The Inverse Gambler's Fallacy The Argument from Design: The Anthropic Principle Applied to Wheeler Universes," *Mind*, LXXVI (1987): 331-40.

[5] For example, Ernan McMullin, "Indifference Principle and Anthropic Principle in Cosmology," *Studies in History and Philosophy of Science*, XXIV, 3 (1993): 359-89

[6] For example, Richard Swinburne, "Argument from the Fine-tuning of the Universe," in Leslie, ed., *Physical Cosmology and Philosophy* (New York: Macmillan, 1990), pp. 154-73

observations are made (which I shall show is impotent as a basis for evaluating Big World theories): we also know that such-and-such observations are made *by us*. This indexical de se component of our evidence turns out to be crucial to cosmology, and recognizing this is the first step to the solution I shall propose.

The second step is to formulate a new methodological principle that describes the probabilistic evidential bearing of (partly) indexical information on nonindexical hypotheses.

With the expanded evidence base and the new rule, we can explain how Big World theories are testable. I shall also hint at how the epistemological theory outlined here is useful in other areas of philosophy and scientific methodology.

But first, let us study in more detail how things go wrong if we construe the evidence nonindexically, in the form 'Such-and-such an observation is made'. We can be generous and take 'an observation' in a broad sense to include the total phenomenological content present in the observer's mind. We do not, however, at this stage take 'observing' as a success verb, implying the veracity of observations; but rather, we assume an internal reading of the evidence. This assumption will later be relaxed.

## I THE CONUNDRUM

Consider a random phenomenon, for instance, Hawking radiation. When black holes evaporate, they do so in a random manner such that, for any given physical object, there is a finite (although extremely small) probability that it will be emitted by any given black hole in a given time interval. Such things as boots, computers, or ecosystems have some finite probability of popping out from a black hole. The same holds true, of course, for human bodies and human brains in particular states.[7] Assuming that mental states supervene on brain states, there is thus a finite probability that a black hole will produce a brain in a state of making any given observation. Some of the observations made by such brains will be illusory and some will be truthful. For example, some brains produced by black holes will have the illusory experience of reading a measurement device that does not exist. Other brains, with the same experiences, will be making veridical observations—a measurement device may materialize together with the brain and may have caused the brain to make the observation. But the point that matters here is that any observation we could make has a finite probability of being produced by any given black hole.

---

[7] See, for example, S. W. Hawking and W. Israel, eds., *General Relativity: An Einstein Centenary Survey* (New York: Cambridge, 1979): "[I]t is possible for a black hole to emit a television set or Charles Darwin" (p. 19). To avoid making a controversial claim about personal identity, Hawking and Israel ought to have weakened this to "an exact replica of Charles Darwin." But see also Gordon J. Belot et al., "The Hawking Information Loss Paradox· The Anatomy of a Controversy," *British Journal for the Philosophy of Science*, L, 2 (1999): 189-229.

The probability of *anything* macroscopic and organized appearing from a black hole is of course minuscule. The probability of a given conscious brain state being created is tinier still. Yet even a low-probability outcome has a high probability of occurring if the random process is repeated often enough. And that is precisely what happens in our world, if the cosmos is very vast. In the limiting case where the cosmos contains an infinite number of black holes, the probability of any given observation being made is one.[8]

There are good grounds for believing that our universe is open or flat and contains an infinite number of black holes. Therefore, we have reason to think that any possible human observation is, in fact, instantiated in the actual world.[9] Evidence for the existence of a multiverse would only add further support to this proposition.

It is not necessary to invoke black holes to make this point. Any random physical phenomenon would do. It seems we do not even have to limit the argument to quantum fluctuations. Classical thermal fluctuations could, presumably, in principle lead to the molecules in a gas cloud containing the right elements to spontaneously bump into each other so as to form a biological structure such as a human brain.

The problem is that it seems impossible to get any empirical evidence that could distinguish between various Big World theories. For any observation we make, *all* such theories assign a probability of one to the hypothesis that that observation is made. That means that the fact that the observation is made is no reason whatever for preferring one of these theories to the others. Experimental results appear totally irrelevant.[10]

---

[8] In fact, there is a probability of unity that infinitely many tokens of each observation type will appear. But one of each suffices for present purposes.

[9] I restrict the assertion to *human* observations in order to avoid questions as to whether there may be other kinds of possible observations that perhaps could have infinite complexity or be of some alien or divine nature that does not supervene on stuff that is emitted from black holes—such stuff is physical and of finite size and energy.

[10] Some cosmologists are recently becoming aware of the problematic that this paper describes—for example, A. Vilenkin, "Unambiguous Probabilities in an Eternally Inflating Universe," *Physical Review Letters*, LXXXI (1998): 5501-04; A. Linde and A. Mezhlumian, "On Regularization Scheme Dependence of Predictions in

We can see this formally as follows. Let $B$ be the proposition that we are in a Big World, defined as one that is big enough and random enough to make it highly probable that every possible human observation is made. Let $T$ be some theory that is compatible with $B$, and let $E$ be some proposition asserting that some specific observation is made. Let $P$ be an epistemic probability function. Bayes's theorem states that

$$P(T|E\&B) = P(E|T\&B)P(T|B)/P(E|B)$$

In order to determine whether $E$ makes a difference to the probability of $T$ (relative to the background assumption $B$), we need to compute the difference $P(T|E\&B) - P(T|B)$. By some simple algebra it is easy to see that

$$P(T|E\&B) - P(T|B) \sim 0 \text{ if and only if } P(E|T\&B) \sim P(E|B)$$

This means that $E$ will fail to give empirical support to $T$ (modulo $B$) if $E$ is about equally probable given $T\&B$ as it is given $B$. We saw above that $P(E|T\&B) \sim P(E|B) \sim 1$. Consequently, whether $E$ is true or false is irrelevant for whether we should believe in $T$, given we know $B$.

To illustrate, let $T_2$ be some perverse permutation of an astrophysical theory $T_1$ that we actually embrace. $T_2$ differs from the $T_1$ by assigning a different value to some physical constant. To be specific, let us suppose that $T_1$ says that the temperature of the cosmic microwave background radiation is about 2.7 Kelvin (which is the observed value), whereas $T_2$ says it is, say, 3.1 K. Suppose furthermore that both $T_1$ and $T_2$ imply that we are living in a Big World. One would have thought that our experimental evidence favors $T_1$ over $T_2$. Yet the above argument seems to show that this view is mistaken. Our observational evidence supports $T_2$ just as much as $T_1$. We really have no reason to think that the background radiation is 2.7 K rather than 3.1 K.

## II. IT IS NOT THE OLD POINT ABOUT UNDERDETERMINATION OF THEORY BY DATA

At first blush, it could seem as if this simply rehashes the lesson, made familiar by Pierre Duhem and W. V. Quine,[11] that it is always possible to rescue a theory from falsification by modifying some auxiliary assumption, so that strictly speaking no scientific theory ever implies

---

Inflationary Cosmology," *Physical Review D*, LIII (1996): 4267-74. See also Leslie, "Time and the Anthropic Principle," *Mind*, CI, 403 (1992): 521-40.

[11] Duhem, *The Aim and Structure of Physical Theory*, Philip P. Wiener, trans. (Princeton: University Press, 1955); Quine, "Two Dogmas of Empiricism," *Philosophical Review*, LX (1951): 20-43.

any observational consequences. The above argument would then merely have provided an illustration of how this general result applies to cosmological theories. But this would be to miss the point.

If the argument given above is correct, it establishes a much more radical conclusion. It purports to show that all Big World theories are not only logically compatible with any observational evidence, but they are also *perfectly probabilistically compatible*. They all give the same conditional probability (namely, one) to every observation statement $E$ defined as above. This entails that no such observation statement can have *any* bearing, whether logical or probabilistic, on whether the theory is true. If that were the case, it would not seem worthwhile to make astronomical observations if what we are interested in is determining which Big World theory to favor. The only reasons we could have for choosing between such theories would be either a priori (simplicity, elegance, and the like) or pragmatic (such as ease of calculation).

Nor is the argument making the ancient statement that human epistemic faculties are fallible, that we can never be certain that we are not dreaming or are brains in a vat. No, the point here is not that such illusions *could* occur, but rather that we have reason to believe that they *do* occur, not just some of them but all possible ones. In other words, we can be fairly confident that the observations we make, along with all possible observations we could make in the future, are being made by brains in vats and by humans that have spontaneously materialized from black holes or from thermal fluctuations. The argument would entail that this abundance of observations makes it impossible to derive distinguishing observational consequences from contemporary cosmological theories.

## III. THE CONCLUSION IS A REDUCTIO

I trust that most readers will find this conclusion unacceptable. Cosmologists certainly appear to be doing experimental work and modify their theories in light of new empirical findings. The COBE satellite, the Hubble Space Telescope, and other devices are showering us with data that have been causing something of a renaissance in the world of astrophysics in recent years. Yet the argument described above would show that the empirical import of this information could never go beyond the humble role of providing support for the hypothesis that we are living in a Big World—for instance, by showing that the universe is open. Nothing apart from this one fact could be learnt from such observations. Once we have established that the universe is open and infinite, then any further work in observational astronomy would be a waste of time and money.

Worse still, the leaky connection between theory and observation in cosmology spills over into other domains. Since nothing hinges on how we defined *T* in the derivation above, the argument can easily be extended to prove that observation does not have a bearing on any scientific question so long as we assume that we are living in a Big World.[12]

This consequence is absurd, so we should look for a way to mend the methodological pipeline and restore the flow of testable observational consequences from Big World theories. How can we do that?

## IV. GIVING UP THE INTERNAL CONSTRUAL OF 'OBSERVATION' DOES NOT SAVE US

Suppose we give up the internal construal of 'observation' and instead take the term as a success verb, so that observing, say, a blue table implies that there is a blue table that is causally responsible for the observation. Suppose further that we couple this with the postulation that we are entitled (and perhaps even required) to have a prior credence function that strongly favors the hypothesis that we for the most part really do observe (in the success sense) what it seems to us that we are observing. Then it might appear as if we have an exit from our predicament. (Alternatively, we could formulate this escape plan by sticking to the original internal definition of 'observation' and adding the postulate that our prior credence functions should strongly favor the veridicality of our observations.)

Even setting aside foundationalist scruples, however, the proposed solution does not get us out of the pickle.

To see this, consider that observers are not the only things that have a finite probability of being generated in random systems. On the same ground that we should expect human observers in all possible states to be ejected from black holes or to form from vastly improbable thermal fluctuations, we should also expect all physically possible local environments to spring forth. So not only are there observers having all sorts of illusions (of seeing a blue table or of reading a measurement apparatus) but additionally there are observers making all sorts of *veridical* observations (actually seeing a blue table or reading off instruments in each of their possible output states). Consequently, even if we assume our observations to be veridical, we are still left with the problem that our current best theories give probability one to the existence of all possible such

---

[12] And were it really true that we have no means of testing Big World theories, then it would not even be clear that the empirical support we currently have for such theories could be maintained. Such theories would seem self-undermining in that they would say of their own evidence, in effect, that it was not to be trusted.

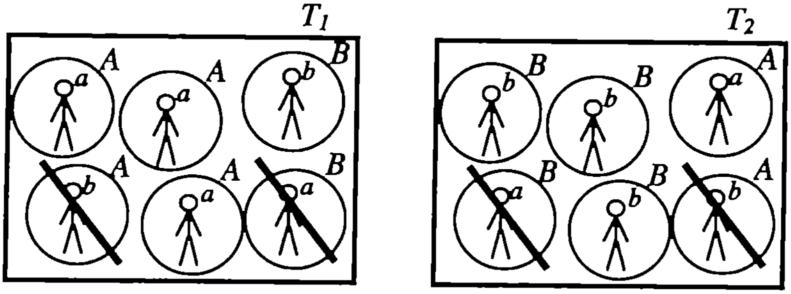


Figure 1: Even if we disregard illusory observations or assume that our observations are veridical, our observation $a$ (seeing the background radiation as 2.7 K) is perfectly compatible with both $T_1$ (which implies that CMB is 2.7 K everywhere ($A$) except where an unusual fluctuation has occurred ($B$)) and $T_2$ (which implies that CMB is generally 3.1 K except for fluctuations).

observations *together with their truth-making local environments.* (See Figure 1.) We can even press on to the conclusion that for any possible human observation, there may be habitats in which that observation is appropriately *caused* by the observed object and in which the observer's perceptions in general track her surroundings.[13]

A qualification is due. While small-scale environments—for example, ones that include tables and measuring apparatuses—are on a par with human bodies, it is not clear that very large systems such as galactic superclusters could be produced by any of the random processes that we have discussed. If we stipulate that we are making veridical observations of these mega-scale entities, we could thus salvage the testability of some aspects of cosmological theories that concern these large-scale entities. Yet this would be of little avail since it would not rescue the rest of our epistemic practices, which deal with medium-sized and small things. Observations of such items would still be subject to the charge of being radically irrelevant to our theories about the world, modulo the Big World hypothesis.[14]

---

[13] I want to emphasize that the problem is not that there is some massive inconsistency of contradictory observations To assert the existence of all possible human observations is not inconsistent, since the observations may be illusory. Moreover, even if all the observations were asserted to be veridical, it would still be no inconsistency, since the various diverse properties that are being observed may be instantiated at different places, just as a tie can be both blue and yellow (although not at the same spot at the same time). Rather, the problem is how to derive testable predictions given our inability to observationally locate ourselves in a Big World, which is rather analogous to seeing a yellow spot through a microscope and not knowing which part of the hypothesized tie we are looking at.

A further shortcoming of the proposal (apart from the fact that it does not work) is that it does not tell us anything about the defeasibility conditions of the purported principle that you should be strongly biased in favor of the veridicality of your observations. Clearly, there are cases where it would be unreasonable to believe that one's observations are veridical. For example, if you knew that almost all observers in your current situation (tucked in, let us say, between the bedsheets in a detox unit with the sensation of bugs crawling under your skin) were hallucinating, then you should *not* believe that your current observations are veridical, unless you had additional information defeating *that* conclusion. A satisfactory account of the Big World case ought to have at least something to say about why the presence of lots of hallucinating and otherwise misled observers in Big Worlds does not undermine our confidence in the reliability of our own observations while the contrary holds specifically for clients in the methadone clinic and other such special situations.

So, if an externalist construal of the evidence is not the answer, what is?

## V. RESTORING THE FLOW OF TESTABLE CONSEQUENCES VIA A LIMITED INDIFFERENCE PRINCIPLE OVER DE SE STATEMENTS

It may seem as if our troubles originate from the somewhat "technical" point that in a large enough cosmos, every observation will be made by *some* freakish observers here and there. It remains the case, however, that those observers are exceedingly rare and far between. For every observation made by a freak observer spontaneously materializing from Hawking radiation or thermal fluctuations, there are trillions upon trillions of observations made by regular observers who have evolved on planets like our own and who make veridical observations of the universe. Why can we not solve the problem, then, by saying that although all these freak observers exist and are suffering from various illusions (or are making veridical but unrepresentative observations), it is highly unlikely that *we* are among their numbers? Then we should think, rather, that we are very probably one of the regular observers whose observations reflect reality. We could safely ignore the freak observers and their illusions and misleading perceptions in most contexts when doing science.

---

[14] We may also note that there are some (speculative) theories according to which even the largest structures that we see are not large enough to escape the problem (for example, M. Tegmark, "Does the Universe in Fact Contain Almost No Information?" *Foundations of Physics Letters*, IX, 1 (1996): 25-42). Moreover, there are many much less extreme theories, such as chaotic inflation theory (see, for example, A. Linde, "Inflation with Variable Omega," *Physics Letters B*, CCCLI (1995): 99-104), according to which observers are observing a wide range of different values of some physical constant and parameters, not because the observers have illusions or live in habitats that originate from black holes or the like, but because the "constants" and parameters vary over vast cosmic distances or epochs.

In my view, this response suggests the right way to proceed. Because the freak observers are in such a tiny minority, their observations can be disregarded for most purposes. It is *possible* that we are freak observers—we should assign to that hypothesis some finite probability, but such a tiny one that it does not make any practical difference.

If we want to run with this idea, it is crucial that we construe our evidence differently than we did above. If our evidence is simply 'Such-and-such an observation is made', then the evidence has probability one given any Big World theory—and we ram our heads straight into the problems I described. But if we construe our evidence in the more specific form '*We* are making such-and-such observations' then we have a way out. For we can then say that although Big World theories make it probable that some such observations be made, they need not make it probable that we should be the ones making them.

Let us therefore define:

$E'$ := "Such-and-such observations are made by us."

$E'$ contains an indexical de se component that the original evidence-statement we considered, $E$, did not. $E'$ is logically stronger than $E$. The rationality requirement that one should take all relevant evidence into account dictates that in case $E'$ leads to different conclusions than does $E$, then it is $E'$ that determines what we ought to believe.

A question that now arises is how to determine the evidential bearing that statements of the form of $E'$ have on cosmological theories. Using Bayes's theorem, we can turn the question around and ask: How do we evaluate $P(E'|T\&B)$, the conditional probability that a Big World theory gives to us making certain observations? The argument in the foregoing sections showed that, if we hope to be able to derive any empirical implications from Big World theories, then $P(E'|T\&B)$ should not generally be set to unity or close to unity. $P(E'|T\&B)$ must take on values that depend on the particular theory and the particular evidence that we are considering. Some theories $T$ are supported by some evidence $E'$; for these choices $P(E'|T\&B)$ is relatively large. For other choices of $E'$ and $T$, the conditional probability will be much smaller.

To be concrete, consider the two rival theories about the temperature of the cosmic microwave background, $T_1$ and $T_2$. Let $E'$ be the proposition that we have made those observations which cosmologists innocently take to support $T_1$. $E'$ includes readings from radio telescopes, and the like. Intuitively, we want $P(E'|T_1\&B) > P(E'|T_2\&B)$. That inequality must be the reason why cosmologists believe that the background radiation is in accordance with $T_1$ rather than $T_2$, since a priori there is no ground for assigning $T_1$ a substantially greater probability than $T_2$.

A natural way to achieve this result is by postulating that we should think of ourselves as being in some sense "random" observers. Here, we use the idea that the essential difference between $T_1$ and $T_2$ is that the *fraction* of observers that would be making observations in agreement with $E'$ is enormously greater on $T_1$ than on $T_2$. If we reason as if we were randomly selected samples from the set of all observers, or from some suitable subset thereof, then we can explicate the conditional probability $P(E'|T\&B)$ in terms of the expected fraction of all observers in the reference class that the conjunction of $T$ and $B$ says would be making the kind of observations that $E'$ says that we are making. As we shall see, this postulate enables us to conclude that $P(E'|T_1\&B) > P(E'|T_2\&B)$.

Let us call this postulate the *self-sampling assumption*:

> (SSA) Observers should reason as if they were a random sample from the set of all observers in their reference class.

The general problem of how to define the reference class is complicated, and I shall not address it here. For the purposes of this discussion, we can think of the reference class as consisting of all observers who will ever have existed. We can also assume a uniform sampling density over this reference class. Moreover, it simplifies things if we set aside complications arising from assigning probabilities over infinite domains by assuming that $B$ entails that the number of observers is finite, albeit such a large finite number that the problems described above obtain. These assumptions help us focus on basic principles.

Here is how SSA supplies the missing link needed to connect theories like $T_1$ and $T_2$ to observation. On $T_2$, the only observers who observe an apparent temperature of the cosmic microwave background CMB $\approx$ 2.7 K are those who either have various sorts of rare illusions (for example, because their brains have been generated by black holes and are therefore not attuned to the world they are living in) or happen to be located in extremely atypical places (where, for example, a thermal fluctuation has led to a locally elevated CMB temperature). On $T_1$, by contrast, almost every observer who makes the appropriate astronomical measurements and is not deluded will observe CMB $\sim$ 2.7 K. A much greater fraction of the observers in the reference class observe CMB $\sim$ 2.7 K if $T_1$ is true than if $T_2$ is true. By SSA, we consider ourselves as random observers; so it follows that on $T_1$ we would be much more likely to find ourselves as one of those observers who observe CMB $\sim$ 2.7 K than we would on $T_2$. Therefore, $P(E'|T_1\&B) > P(E'|T_2\&B)$. Supposing that the prior probabilities of $T_1$ and $T_2$ are roughly the same, $P(T_1) \sim P(T_2)$, it is then trivial to derive via Bayes's theorem that $P(T_1|E'\&B) > P(T_2|E'\&B)$. This vindicates the intuitive view that we do have empirical evidence that favors $T_1$ over $T_2$.

The job that SSA is doing in this derivation is to enable the step from a proposition about fractions of observers to propositions about corresponding probabilities. We get the propositions about fractions of observers by analyzing $T_1$ and $T_2$ and combining them with relevant background information $B$; from this we conclude that there would be an extremely small fraction of observers observing CMB $\sim$ 2.7 K given $T_2$ and a much larger fraction given $T_1$. We then consider the evidence $E'$, which is that *we* are observing CMB $\sim$ 2.7 K. SSA authorizes us to think of the 'we' as a kind of random variable ranging over the class of actual observers. From this it then follows that $E'$ is more probable given $T_1$ than given $T_2$. But without assuming SSA, all we can say is that a greater fraction of observers observe CMB $\sim$ 2.7 K if $T_1$ is true, and at that point the argument would grind to a halt. We could not reach the conclusion that $T_1$ is supported over $T_2$. For this reason I propose that SSA, or something like it, be adopted as a methodological principle.

It may seem mysterious how probabilities of this sort can exist— How can we possibly make sense of the idea that there was some chance that we might have been other observers than we are? What I am suggesting here, however, is not the existence of some objective, or physical, chances. I am not suggesting that there is a physical randomization mechanism, a cosmic fortune wheel as it were, that assigns souls to bodies in a stochastic manner. Rather, we should think of these probabilities as *epistemic*. They are part of a proposal explicating the epistemic relations that hold between theories (such as $T_1$ and $T_2$) and evidence (such as $E'$) containing a de se component. We can view SSA as a kind of restricted indifference principle that applies to credences over de se propositions, or sets of centered possible worlds in the Quinean terminology. The status of SSA could also be regarded as in some respects akin to that of the David Lewis's[15] "principal principle," which expresses a connection between physical chance and epistemic credence. Crudely put, the principal principle says that, if you know that the objective (physical) chance of some outcome $A$ is $x\%$, then you should assign a credence of $x\%$ to $A$ (unless you have additional "inadmissible" information). Analogously, SSA can be read as saying that, if you know that a fraction $x\%$ of all observers in your reference class are in some type of position $A$, then you should assign a prior credence of $x\%$ to being in a type-$A$ position. This prior credence must, of course, be conditionalized on any other relevant information you have in order to get the posterior credence, that is, the degrees of belief you should actually have given all you know. Thus, after conditionalizing on the observation that CMB ~ 2.7 K, you get, trivially, a posterior function that assigns zero credence to the hypothesis that you are an observer who observes CMB ~ 3.1 K. But it is the higher conditional *prior* credence (according to SSA) of observing that CMB ~ 2.7 K given $T_1$ than given $T_2$ that renders it the case that conditionalizing on this observation preferentially supports $T_1$.

## VI. AN ILLUSTRATION

We can illustrate how SSA works by a simple thought experiment.

> *Blackbeards and Whitebeards.*
> In an otherwise empty world there are three rooms. God tosses a fair coin and creates three observers as a result, placing them in different rooms. If the coin falls heads, He creates two observers with black beards and one with a white beard. If it falls tails, it is the other way around: He creates two whitebeards and one blackbeard. All observers are aware of these conditions. There is a mirror in each room, so observers know the color of their own beard. You find yourself in one of the rooms and you see that you have a black beard. What credence should you give to the hypothesis that the coin fell heads?

The situation is depicted in Figure 2.

Because of the direct analogy to the cosmology case, we know that the answer must be that you should assign a greater credence to *Heads* than to *Tails*. Let us apply SSA and see how we get this result.

From the setup, we know that the prior probability of *Heads* is 50%. This is the probability you should assign to *Heads* before you have

---

[15] *Philosophical Papers, Volume II* (New York: Oxford, 1986), and "Humean Supervenience Debugged," *Mind*, CIII, 412 (1994): 473-90. A similar principle had earlier been introduced by Hugh Mellor in *The Matter of Chance* (New York: Cambridge, 1971).

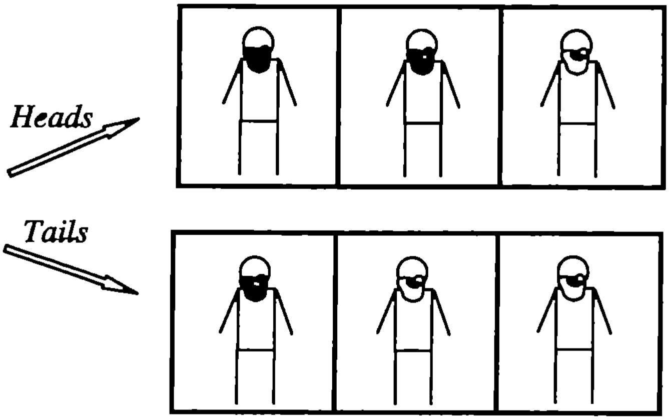


Figure 2: The "Blackbeards and Whitebeards" thought experiment.

looked in the mirror and thus before you know your beard color. That this probability is 50% follows from the principal principle together with the fact you know that the coin toss was fair. We thus have

$$P(Heads) = P(Tails) = \tfrac{1}{2}$$

Next we consider the conditional probability of you observing that you have a black beard given a specific outcome of the toss. If the coin fell heads, then two out of three observers observe themselves having a black beard. If the coin fell tails, then one out of three observe having a black beard. By SSA, you reason as if you were a randomly sampled observer, giving

$$P(Black|Heads) = \tfrac{2}{3} \qquad P(Black|Tails) = \tfrac{1}{3}$$

Using Bayes's theorem, we can then calculate the conditional probability of *Heads* given that you have a black beard:

$$P(Heads|Black) = \frac{P(Black|Heads)P(Heads)}{P(Black|Heads)P(Heads) + P(Black|Tails)P(Tails)}$$

$$= \frac{\tfrac{2}{3} \cdot \tfrac{1}{2}}{\tfrac{2}{3} \cdot \tfrac{1}{2} + \tfrac{1}{3} \cdot \tfrac{1}{2}} = \tfrac{2}{3}$$

After looking in the mirror and learning that your beard is black, you should therefore assign a credence of ⅔ to *Heads* and ⅓ to *Tails.*

This result mirrors that of the cosmology example. Because one theory ($T_1$, *Heads*) entails that a greater fraction of all observers are observing what you are observing ($E'$, *Black*) than does another theory ($T_2$, *Tails*), the former theory obtains preferential support from your observation.

## VII. SUMMARY WE NEED A METHODOLOGY FOR EVIDENCE WITH A DE SE COMPONENT

Big World theories, popular in contemporary cosmology, engender a peculiar methodological problem: because they say the world is very big and somewhat stochastic, they imply (or make it highly probable) that every possible human observation is made. The difficulty is that it is unclear how we could ever have empirical reasons for preferring one such theory to another, since they all seem to fit equally well with whatever we observe. This skeptical threat is different from and much more radical than the problem of underdetermination of theory by data associated with Duhem and Quine. And if left unfixed, the broken connection between observation and theory spills over from cosmology into other domains.

We saw that the leakage cannot be stopped even by blocking all consideration given to the possibility of illusory observations, because the maverick observations made in Big Worlds include veridical ones as well as illusions. Instead, we proposed to repair methodology by means of a new epistemic principle, the self-sampling assumption, which takes into account the de se component of our evidence. This principle connects Big World theories to observation in an intuitively plausible way and vindicates the practices of cosmologists who test hypotheses against experimental findings.

The self-sampling assumption has implications in other problem areas in science and philosophy. It can be seen as an explication of the anthropic principle, understood in the original spirit of Brandon Carter,[16] a theoretical physicist whose seminal work opened the door to a systematic exploration of observation selection effects. Observation selection effects are a kind of bias which may be present in our data which is not due to limitations in our measurement apparatuses but to the fact that our data are preconditioned on the existence of a suitably positioned observer to "have" the data (and to build the instruments in the first place). Carter investigated the relevance of observation selection effects for attempts to evaluate the bearing of our current evidence on questions such as how improbable it is for complex life forms to evolve on a given Earth-like planet or how many critical improbable steps were involved in our evolution.[17] To illustrate, take one of the simplest points Carter made: even if a theory says that the probability for an Earth-like planet giving rise to intelligent life is small, the theory will still perfectly fit our observation of intelligent life having evolved on this planet provided that the total number of Earth-like planets is large enough for it to have been probable, according to the theory, that intelligent life should arise somewhere.

---

[16] "Large Number Coincidences and the Anthropic Principle in Cosmology," in M. S. Longair, ed., *Confrontation of Cosmological Theories with Data* (Boston: Reidel, 1973), pp. 291-98; "The Anthropic Selection Principle and the Ultra-Darwinian Synthesis," in F. Bertola and U. Curi, eds., *The Anthropic Principle* (New York: Cambridge, 1989), pp. 33-63.

Similar modes of reasoning are invoked in some discussions of no-collapse versions of quantum mechanics[18] and, as hinted at in the introduction, they play a central role in the debate about the significance of the apparent fine-tuning of our universe and the capacity of multiverse theories to explain it. Even an application to traffic planning has been discovered.[19] On the more theoretical side, we have game theoretic problems involving imperfect recall, such as the Absent-minded Driver problem[20] and its philosophical, more purely epistemic analogue, the Sleeping Beauty problem.[21]

What these various topics have in common is that they involve the assignment of conditional credences to statements of the form 'I make such-and-such observations given that the world is such and such'.[22] In other words, they involve the evaluation of a de se component of our evidence: our knowledge that *we* are the ones making a certain observation or that *we* are the ones who have a certain piece of (otherwise nonindexical) evidence. Our duty to objectivity must not be misunderstood as a license to ignore de se clues. The considerations advanced here impose constraints on what can count as a satisfactory methodology for fashioning knowledge out of this indexical part our epistemic raw material. Such a methodology, a general theory of observation selection effects and its various scientific and philosophical applications, is something I have attempted to set forth elsewhere.[23]

---

[17] "The Anthropic Principle and Its Implications for Biological Evolution," *Philosophical Transactions of the Royal Society A*, CCCX (1983): 347-63.

[18] See, for example, D. N. Page, "Can Quantum Cosmology Give Observational Consequences of Many-Worlds Quantum Theory?" in C. P. Burgess and R. C. Myers, eds., *General Relativity and Relativistic Astrophysics, Eighth Canadian Conference, Montreal, Quebec* (New York: American Institute of Physics, 1999), pp. 225-32.

[19] My "Cars in the Next Lane Really Do Go Faster," *PLUS*, XVII (2001).

[20] See, for example, Michele Piccione and Ariel Rubinstein, "On the Interpretation of Decision Problems with Imperfect Recall," *Games and Economic Behaviour*, XX (1997): 3-24; Robert J. Aumann, Sergiu Hart, and Motty Perry, "The Forgetful Passenger," *Games and Economic Behaviour*, XX (1997): 117-20.

[21] For example, Adam Elga, "Self-locating Belief and the Sleeping-Beauty Problem," *Analysis*, LX, 266 (2001): 143-47; Lewis, "Sleeping Beauty: Reply to Elga," *Analysis*, LXI, 271 (2001): 171-75.

[22] Or in some cases, the analogous temporal construction: "I make such-and-such observations *now* given that the world is such and such."

NICK BOSTROM

Oxford University

---

[23] *Anthropic Bias. Observation Selection Effects in Science and Philosophy* (New York: Routledge, 2002). A theory of observation selection effects must walk a fine line in order to cater to legitimate scientific needs while avoiding philosophical paradoxes, of which a great number lie in ambush. Incidentally, the self-sampling assumption is, in my view, a mere derivative of a more powerful principle, and it is only valid in special cases.

# BELIEF AND TRANSLATION[*]

I intend to advocate a reform of our pretheoretical intuitions concerning belief ascriptions. In particular, I shall try to convince the reader that it is incorrect to describe a speaker who sincerely assents to 'Cicero is bald' while rejecting 'Tully is bald' as believing that Cicero is bald but Tully is not. In doing so, I shall sketch a theory of belief ascriptions which validates both quantifying into belief contexts (provided the relevant that-clause involves only proper names and, possibly, other devices of direct reference) and an amended form of Leibniz's law. My starting point will be Saul Kripke's puzzle about belief, which not only is the central problem with which any theory of belief ascription must grapple, but which also reveals the close connection between belief ascription and translation. The theory of belief ascriptions presented here exploits this basic insight and, incidentally, does not resort to the usual apparatus of senses, concepts, and propositions.

## I

A powerful result is implicit in Kripke's "A Puzzle about Belief,"[1] where he shows that in some circumstances, no translation could possibly exist between two languages, one of which is powerful enough to ascribe beliefs to a competent and sincere speaker of the other. I shall first state the result and then draw on some of its remarkable consequences. By way of simplification, suppose that Pierre, instead of being a bilingual speaker—who first asserts in French 'Londres est une jolie ville' and then, in English, 'London is not pretty'—speaks only one language superficially resembling English (but not to be confused with it). Let this be the only language in which he is competent. In this language—call it $L_{Pierre}$—two names exist that refer to London: 'London' and 'Londres'.[2] Let $A_{Pierre}$ be the set of statements to which he has a disposition sincerely and reflectively to assent. And let both 'Londres is pretty' and 'London is not

---

[*] Special thanks to Paolo Casalegno and Gennaro Chierchia for their aid and encouragement. Thanks also to the participants at the David Kaplan Conference, "From This and That to Ouch and Oops," San Marino (May 2001), and to those at the Summer School in Analytic Philosophy, Parma, Italy (July 2001).

[1] In Avishai Margalit, ed., *Meaning and Use* (Boston: Reidel, 1979), pp. 239-83, reprinted in Nathan Salmon and Scott Soames, eds., *Propositions and Attitudes* (New York: Oxford, 1988), pp. 102-48

[2] Kripke himself mentions this form of the puzzle. He cites the case of Hebrew, in which there are two names for 'Germany', namely, 'Germaniah' and 'Ashkenaz'.

0022-362X/02/9912/624–47

pretty' belong to $A_{Pierre}$. As I mean to show that every form of the puzzle crucially involves translation, even in the case of monolingual Peter and his uses of the name 'Paderewski', let us distinguish Pierre's language as clearly as possible from our own. I shall here take advantage of the simplifying assumption that Pierre is competent in only one language, $L_{Pierre}$. Let $L_{Us}$ be our language and $T_{Us}$ be the set of all true statements of $L_{Us}$ of the form 'Pierre believes that ...'. Let $M$ be the set of admissible translation functions from $L_{Pierre}$ to $L_{Us}$. Let us assume—though this is *not* in Kripke's original article—that all $\boldsymbol{m}$s in $M$ are compositional. As far as the logical constants are concerned, this amounts to requiring that $\boldsymbol{m}(\neg p) = \neg \boldsymbol{m}(p)$, $\boldsymbol{m}(p \wedge q) = \boldsymbol{m}(p) \wedge \boldsymbol{m}(q)$, and so on for the other connectives. In the proof of Kripke's result, we shall make use of only the clause concerning negation, but the full strength of the compositionality assumption will be exploited later on. The assumptions explicitly listed by Kripke in deriving his paradoxical conclusion are, first, the principle of translation, to the effect that, if a sentence of one language expresses a truth in that language, then any translation of it into any other language also expresses a truth in that other language. To this, we now add the requirement that reference is also to be preserved, in the sense that, if a term of one language refers to an object, then any translation of it into any other language refers to the same object. Applied to Pierre's case, we can express this more formally as follows:

> *The translation principle* (TP): for every $\boldsymbol{m}$ and for every $p$, if $p$ is true in $L_{Pierre}$, then $\boldsymbol{m}(p)$ is true in $L_{Us}$. In addition, for every term $t$, $t$ and $\boldsymbol{m}(t)$ co-refer.

The disquotation principle says that if a normal speaker, on reflection, sincerely assents to '$p$', then he believes that $p$—with the simplifying assumption that the sentence replacing '$p$' lacks indexical or pronominal devices or ambiguities. It is formally expressed (applied to Pierre's case) as follows:

> *The weak disquotation principle* (WDP): for every $\boldsymbol{m}$ and for every $p$, if $p$ belongs to $A_{Pierre}$, then 'Pierre believes that $\boldsymbol{m}(p)$' belongs to $T_{Us}$.

Finally, the principle of consistency, not explicitly stated, but used by Kripke, is expressed (as applied to Pierre) as follows:

> *The consistency principle* (CP): for every $s$ in $L_{Us}$, if 'Pierre believes that $s$' belongs to $T_{Us}$, then 'Pierre believes that $\neg s$' does not belong to $T_{Us}$.

From this, by WDP, it follows in particular that for every $\boldsymbol{m}$ in $M$, there is no $p$ or $q$ in $A_{Pierre}$ such that $\boldsymbol{m}(p)$ and $\boldsymbol{m}(q)$ are of the form $s$ and $\neg s$, respectively.

Here is the puzzle. Choose some particular translation function, $m_0$, if one exists. For simplicity's sake, we may assume without any loss of generality that for any expression $E$, predicate or proper name other than 'London' and 'Londres' in $L_{Pierre}$, $m_0(E) = E$ (as an expression in $L_{Us}$).[3] We reason as follows. Pierre assents to both 'Londres is pretty' and 'London is not pretty', which therefore both belong to $A_{Pierre}$. By WDP, 'Pierre believes that $m_0$('Londres is pretty')' and 'Pierre believes that $m_0$('London is not pretty')' are both in $T_{Us}$. Given that, (a), Kripke's story makes it clear that both of Pierre's 'London' and 'Londres' refer to London, (b), that London has only one name in English, 'London', and, (c), that $m_0$ is compositional, by TP we conclude that both 'Pierre believes that London is pretty' and 'Pierre believes that London is not pretty' are in $T_{Us}$. This goes against CP.

The puzzle has been taken by various authors as showing that either WDP or CP is untenable. Tyler Burge states that the WDP is "absurd"; Ruth Barcan Marcus and Robert Brandom[4] take similar stances. Some philosophers who advocate a Russellian view of belief oppose CP forcefully. Nathan Salmon,[5] for one, claims: "In hindsight,..., [Kripke's] examples and arguments are today very often seen—perhaps even usually seen—as making an extremely strong case (however inadvertent) against [Pierre's] Consistency [that is, $\neg$([*Pierre*] *believes* $p \wedge$ [*Pierre*] *believes* $\neg p$) ]", even though he acknowledges that Kripke himself "adamantly opposes this conclusion in particular" (*ibid.*, p. 5). The idea here is that Pierre can consistently believe both that $p$ and that $\neg p$ at the same time, provided that he does so under two different modes of presentation, without realizing that in both cases he is focusing on the same proposition. This, of course, is incompatible with the *strong disquotation principle* (SDP), namely, the assumption that, for every $m$, $p$, and $s$, if $p$ belongs to $A_{Pierre}$, then 'Pierre believes that $m(p)$' belongs to $T_{Us}$, and, conversely, if 'Pierre believes that $s$' belongs to $T_{Us}$, then for some (compositional) translation $n$ from our language into Pierre's language, $n(s)$ belongs to $A_{Pierre}$. If Pierre, against consistency, can be said to believe both that $s$ and that $\neg s$, for some $s$, then in virtue of SDP he ought to have a disposition to assent both to $n(s)$ and to $n(\neg s)$, that is, $\neg n(s)$—which would make Pierre overtly incoherent. According to other philosophers, the moral to be drawn by Pierre's puzzle is precisely that translation may fail to be compositional or just that it may fail to commute with negation. I am not interested in either criticizing or defending the interpretations of the puzzle that have been given so far. In the end, they are as good as the theories of belief ascription which they help to build up. But it has to be pointed out that, in particular, giving up compositionality is no minor loss. As a matter of fact, compositionality of translation will play a rather large role in the second part of the present discussion. It will be appealed to in accounting for some important phenomena (for example, that existential generalization fails to preserve truth in general) that are hard to explain otherwise. Even though I do not have direct arguments to support the compositionality hypothesis, I find this indirect evidence for it rather compelling. I shall come back to this issue in a subsequent section.

---

[3] Here, and below, I let myself be somewhat lax in matters of use and mention. Clearly, translation functions take expressions in one language to expressions of another (or the same) language

[4] Burge, "Two Thought Experiments Revisited," *Notre Dame Journal of Formal Logic*, XXIII (1982): 284-93; Marcus, "Rationality and Believing the Impossible," in her *Modalities: Philosophical Essays* (New York: Oxford, 1993), pp. 144-61; Brandom, *Making It Explicit* (Cambridge: Harvard, 1994).

[5] "Being of Two Minds," *Noûs*, XXIX, 1 (1995): 1-20.

As I said, I shall not discuss here the interpretations of the puzzle. But it seems to have escaped notice that beyond the principles explicitly listed above, there is the crucial assumption that the set of translations $M$, from $L_{Pierre}$ to $L_{Us}$, is nonempty. Accordingly, the brunt of the argument can be turned against it: the puzzle can be taken as containing a perfectly cogent proof that, in some circumstances, no compositional translation exists between two languages, one of which is sufficiently rich for ascribing beliefs to a speaker of the other, and has sufficiently many proper names (not to mention other kinds of singular terms and other resources) to name each one of the entities acknowledged by the other.

This is an impossibility result. Let us state it in the full generality allowed us by the apparatus introduced so far.

> *Kripke's result*: there is no translation $m$ from $L_{Pierre}$ to $L_{Us}$ satisfying the principles TP, WDP, and CP, which is such that for some two terms $t_1$ and $t_2$ in $L_{Pierre}$, $m(t_1) = m(t_2)$, and for some sentence $p$, both $p[t_1]$ and $\neg p[t_2]$ are in $A_{Pierre}$.

We can think of 'London' and 'Londres' as playing the role of $t_1$ and $t_2$, for we seem to have no other choice but to translate them both as 'London' in accordance with TP and because English has no other name for the capital of England. Here, '$x$ is pretty' plays the role of $p(x)$. The proof of the result is straightforward.

> *Proof.* If, for some $p$, both $p[t_1]$ and $\neg p[t_2]$ are in $A_{Pierre}$—that is, if Pierre assents to them—then, by WDP, 'Pierre believes that $m(p[t_1])$' and 'Pierre believes that $m(\neg p[t_2])$' are both in $T_{Us}$. Since $m(t_1) = m(t_2)$ and

> $m$ is assumed to be compositional, then, for some $s$, 'Pierre believes that $s$' and 'Pierre believes that $\neg s$' are both in $T_U$, contrary to CP. If '$t_1 \neq t_2$' is in $A_{Pierre}$— think of Pierre as assenting to 'London ≠ Londres'— then, as presumably everyone must assent to any identity of the form '$t = t$', the same conclusion follows.

Note that nothing in the proof depends on $t_1$ and $t_2$ being proper names or any other particular sorts of singular terms, simple or complex: all that matters is whether or not they are mapped by a translation $m$ onto the same term belonging to the language in which belief is ascribed. The result can be generalized to cover also the case of two predicates that the believer takes to be noncoextensive, even though the translation maps them both onto the same predicate in the ascriber's language. The generalization does not require any assumption not already appealed to in the less general form stated above. As a matter of fact, Kripke himself pointed out (in his original 1979 paper) how general his puzzle was: "paradoxes analogous to the one about London obviously can arise for rabbits and beeches."[6] I suspect that not much is to be obtained from this generalization of the result, however. Intuitively, the special case of the puzzle involving proper names is more compelling, if only because failures to realize that two predicates are coextensive can sometimes (but not always, of course) be blamed on the speaker's linguistic incompetence, and this has a tendency to make the main point appear less dramatic. Moreover, even though the consequences of the result go well beyond the issues surrounding proper names and other singular terms, as we shall see, only its restricted form stated above will be needed for the arguments to follow.

That the result is already contained in Kripke's original article is somewhat obscured by the fact that Pierre is described there as speaking two distinct languages—French and English—each of which is separately translatable into the language in which the ascriptions of Pierre's beliefs are couched, that is, English. This might be confusing, but the result also holds for this slightly more general case involving several languages: it amounts to a denial that a *simultaneous* translation exists from those languages into another (or the same) language. Nothing much is gained, however, by such generalization and I prefer to stick to the simpler and more perspicuous version of the result outlined above.

---

[6] More accurately: "In 'Naming and Necessity' I suggested that there are important analogies between proper names and natural kind terms, and it seems to me that the present puzzle is one instance where the analogy will hold. Putnam, who has proposed views on natural kinds similar to my own in many respects, stressed this extension of the puzzle in his comments at the Conference" (*op. cit.*, p. 264).

Long ago, Quine showed that, if one translation exists between two languages, then there are indefinitely many nonequivalent ones. Belief ascriptions were conspicuously absent from Quine's arguments. Now, taking belief into account, Kripke's result shows that not even one translation exists between some languages that are equally expressive, in the sense that they share, in addition to their logical apparatus, the same stock of predicates, and each contains at least one proper name referring to any entity that can be named in the other language. This is not to deny, of course, that for most practical purposes it is good enough to translate a proper name such as 'Londres' by some such definite description as 'the city called "Londres" '. But even apart from the general issue of the nonequivalence of proper names and definite descriptions, note that translating a proper name by means of a co-referring definite description infringes upon the requirement that translation be compositional.

It might be objected at this point that I am putting too much emphasis on translation, given that one form of Kripke's puzzle—the Paderewski case—does without translation. The case is similar to that of Pierre and 'London'/'Londres'. Peter hears one day of a famous musician, Paderewski, and thinks to himself 'Paderewski has musical talent'. Later on, he is introduced to the famous Polish statesman Paderewski. Without realizing that he is the same person as the famous musician—thinking that politicians are poor musicians—he thinks to himself 'Paderewski does not have musical talent'. Once again, the puzzle is: Does Peter believe, or does he not, that Paderewski has musical talent? Contrary to what happens with Pierre, whose beliefs in the original version of the puzzle are expressed partly in French and partly in English, no translation seems to be involved here. To think so, however, is a mistake. Peter is assumed to be a competent speaker of his language, whatever language it is. Now, a competent speaker is certainly required to be able to tell whether a string of marks or sounds expresses in his language one or more than one word. Two speakers of the same language disagreeing as to how many homophonous words are expressed by a particular string cannot be both equally competent. In Italian, for instance, 'Monaco' is both the name of a city in Germany English speakers call 'Mūnich', and of a small independent state surrounded by France on the Mediterranean, in addition to meaning the same as the English word 'monk'. No one ignorant of the fact that in Italian '[M]onaco' can be both a common noun and a proper name (which, among other things, differ in the syntactic category to which they belong), can be said to be fully competent in Italian. Of course, in order to conclude that we have here three distinct words, rather than one or two, it is enough to note that their referents are distinct. Now, if Peter's language were the same as ours, then the fact that he believes that there are two Paderewskis and therefore two distinct proper names that are spelled the same, whereas we know that there is only one person and one name, would run against the assumption that Peter is a fully competent speaker of his language. It might be objected that it is not required of competent speakers of any language to know how many persons bear any given proper name—or, as I would prefer to put it, how many proper names are expressed by the same string of marks or sounds. Now, it is not obvious to me that this is so, for it certainly *is* required of competent speakers to be able to discern a proper name from a common noun that is spelled the same (as in the case of 'Monaco' which refers, in Italian, both to the city of Mūnich and to the members of a kind of religious community). What are the reasons for maintaining that it is otherwise with two proper names (again, as in the case of 'Monaco' which names, in Italian, both Mūnich and the city where the Monte Carlo gambling casino is located)? This is precisely one of those cases in which the distinction between dictionaries and encyclopedias is not easily drawn. The objection thus fails to be conclusive. Moreover, there is another reason for maintaining that Peter's language is not the same as ours. Suppose it were otherwise. In our language, the relevant identity 'Paderewski = Paderewski' has the form $a = a$. Obviously, Peter must be taking it to be of the form $a = b$, for he thinks that there are two distinct, though homophonous, names involved in it. This conflicts with the assumption that Peter is a fully competent speaker of our language—which I take in the sense that, whatever his encyclopedic knowledge, he makes no mistakes as to the *syntactic* structure of the sentences at hand, for how could two speakers disagreeing as to the syntax of any given sentence be both fully competent in one and the same language? If we maintain the assumption that Peter is fully competent in his own language, therefore, we must conclude that his language is distinct from our own. Some translation is thus needed.

Let us now consider some of the consequences of Kripke's result. There is a naïve, albeit extremely plausible, theory of belief ascription that goes as follows:

> *The naïve theory* (NT): '$S$ believes that $p$' is true relative to some context $C$ (and assignment $f$) if and only if $S$ is disposed to assent to some sentence '$p^*$', in a language he knows, as uttered in the same context $C$ and relative to $f$, which has the same semantic content as '$p$'.

The notion of *sameness of content* can be spelled out in a number of ways—for example, as intensional isomorphism, as same-saying, and as sameness of the proposition expressed (for various notions of proposition)—yielding a family of well-known theories. But, of course, there is one further, rather obvious, possibility, namely, intertranslatability. (I shall ignore the complications arising from the fact that some translation might not be reversible.[7]) It is this theory, supplemented with TP and CP—each independently plausible—that Kripke's result proves to be untenable. Now, how is NT to be amended in light of this result? Since NT is so powerfully supported by intuition, it is advisable that the change be kept to a minimum. According to our diagnosis, our troubles with Pierre stem from the fact that no translation is available between his language and ours. We seem to have no trouble in ascribing beliefs when a suitable translation is available, but we are definitely at a loss when no translation is at hand. If this is what the puzzle shows, then in order to amend the theory, the remedy must surely consist in simply restricting its scope to those cases where a translation *is* available between the language of the believer and that in which the ascriptions are framed. We need not tinker with the notion of translation or resort to sophisticated notions of content. The amended version of the NT then runs as follows:

> *The amended theory* (AT): if there is a translation between the language of the believer, *S*, and our own (or, more precisely, between fragments of such languages) which satisfies TP and CP, then '*S* believes that *p*' is true relative to some context *C* (and assignment *f*) if and only if *S* is disposed to assent to some sentence '*p**' (in *S*'s own language, as uttered in context *C* and relative to *f*) that is intertranslatable with '*p*'. Otherwise—that is, if no translation exists that does justice to the subject's rationality while faithfully taking into account what he asserts or assents to—then ascriptions of the form '*S* believes that...' are neither true nor false.

In the case of Pierre, I find it intuitively plausible that no true belief ascription of the form 'Pierre believes that...is pretty' is available in our language, even before we reach Kripke's result. No reform of intuition is needed so far. Kripke's result only provides the rationale, on quite general grounds, for trusting our pre-existing intuition and AT. Incidentally, AT also says that believing is not the same as assenting to some sentence or other, or standing in some relation with some sentence or other, as Pierre does assent to both 'Londres is pretty' and 'London is not pretty', though he believes neither—the reason being precisely that, in assenting to either of them, he also assents to the other.[8] It follows that while assent is an entirely local matter, involving each sentence separately, belief is not: in order to ascribe a belief to a person, all the sentences to which that person assents, as well as all those she rejects, must be taken into account.[9] This, of course, is *not* to say that we cannot ascribe any belief to Pierre, or to any other believer in the same position as Pierre. We can, but we must be careful about which language we are using in so doing. The obvious course is to resort to Pierre's own language: no impossibility result holds in this case, as the identity mapping is trivially available as a perfectly acceptable translation. This might account, in part, for the rather special status of belief self-ascriptions as compared to ascriptions to others.

---

[7] This issue is itself quite interesting. Kripke's result involves only translations from the language of the believer into that in which belief ascriptions are couched, and no other translation is considered in the main text. Some theories, such as the one defended in Mark Richard, *Propositional Attitudes* (New York: Cambridge, 1990), take the relevant translations to go in the reverse direction. In "Reforming Intuitions about Belief," in Marconi, ed., *Knowledge and Meaning* (Vercelli: Mercurio, 2000), pp. 85-112, I pointed out some of the unwelcome consequences of this.

Here is the provisional conclusion of the argument presented so far. 'Pierre believes that London is pretty' is neither true nor false as a sentence of our own language or idiolect but 'Pierre believes that London is not pretty, while Londres is' is true as a sentence of Pierre's language or idiolect. Alternatively, we can add to our own language some extra foreign name, such as 'Londres'. This is enough both to make Pierre's language translatable into ours, and to make his case entirely parallel to that of our old acquaintance Ralph, who assented to the statement 'Cicero is bald' while dissenting to 'Tully is bald'. It is to be noted that Kripke remarks that "complete and straightforward descriptions of the situation are possible and in this sense there is no paradox" (*op. cit.*, p. 259). Resorting to Pierre's own language gives us one such true and complete description of the situation. This is not yet the solution to the paradox, however. A solution (or, rather, a first step toward the solution, as we shall see in a while) consists in pointing out that every belief ascription is, not so much relative to, but couched in some language or other and that more than one language is involved in the paradoxical cases. The first thing that the paradox urges us to do is to make explicit which language we are using. That the difference between Pierre's language and our own is so very slight, amounting to a mere addition of one proper name, might mislead us into underestimating the significance of adopting the believer's own language. In general, ascribing beliefs to someone else consists, in W. V. Quine's[10] famous phrase, in "projecting ourselves into what, from his remarks and other indications, we imagine [his] state of mind to have been" (*ibid.*, p. 219). Pierre's case shows that our language may prove to be an obstacle to this, as there is no sentence in it, of the form 'Pierre believes that...', which is both intuitively true and a faithful description of his state of mind. Stepping out of our language and into Pierre's own, therefore, amounts to adopting a viewpoint that was simply not available before.

---

[8] Confusion on this score might have originated some sentential theories of belief ascription.

[9] This accounts for the theoretical character of belief, which was clearly perceived by Carnap, among others; see, in particular, his *Meaning and Necessity* (Chicago: University Press, 1956), Supplement C, "On Belief Sentences," pp. 230-32.

## II

The position we have reached so far is not uncomfortable. Making the dependence of belief ascriptions on the language employed explicit seems to defuse the threat posed by Kripke's puzzle to any theory of belief ascription. It is, however, an unstable position, as the reasoning leading to Kripke's result applies—as we shall presently see—not just to cases like Pierre's, but also to cases in which the languages involved are not short of proper names to refer to the relevant items in the world. A rather radical reform of our intuitions concerning belief will then appear to be needed and Pierre's puzzle will prove to be quite pivotal in the issue of belief ascription. Even the provisional conclusion reached above will have to be corrected, at least in part.

Let us reconsider Ralph, who assents to 'Cicero is bald' while rejecting 'Tully is bald'. There is no shortage of names in our language to express what Ralph believes. But is it correct to describe him as believing that Cicero is bald and Tully is not? This is, of course, what Gottlob Frege assumed, and what most people hold, as far as I know. There are some exceptions—notably in Russellian quarters—but even among them the basic intuition seems to point in that direction. Suppose that Frege is right. We then have:

(1) Ralph believes that Cicero is bald.
(2) Ralph does not believe that Tully is bald.

Now, a number of familiar questions are inescapable, among which the following simple one: Can we infer from (1) and (2), respectively:

(3) There is an $x$ such that Ralph believes that $x$ is bald.
(4) There is an $x$ such that Ralph does not believe that $x$ is bald.

---

[10] *Word and Object* (Cambridge: MIT, 1960).

If we can, then we should also have:

(5) There is an $x$ such that $x$ = Cicero and Ralph believes that $x$ is bald.
(6) There is an $x$ such that $x$ = Tully and Ralph does not believe that $x$ is bald.

If quantifying in retains its standard meaning and has its orthodox semantics, however, the truth of both (5) and (6) clashes with the identity 'Cicero = Tully'. More generally, if we simultaneously hold

(a) that $\xi = \zeta$,
(b) that both
  ($b_1$) $S$ believes that...$\xi$...
  ($b_2$) $S$ does not believe that...$\zeta$...
  are true,
(c) that existential generalization is truth preserving, and
(d) that quantifying in (with its orthodox semantics) makes sense

then we run into a straightforward contradiction.[11] How are we to account for this? Some philosophers[12] hold that existentially generalizing from (1) to (3)—the positive case—is fine, but it is not from (2) to (4)—the negative case. This is hardly tenable, however, as the two cases are entirely parallel, and there is no reason for invidiously regarding one transition as truth preserving while rejecting the other.

It is generally agreed that going from de dicto to de re ascriptions is *not* in general permissible. Now, (1) and (2) could very well be taken to be de dicto—though there is room here for disagreement—while (3) and (4) surely count as being de re on any of the various characterizations available of the de dicto/de re divide. That the step from de dicto to de re fails in general is indisputable. Clearly, 'Ralph believes that some dogs have fleas' does not imply 'There are some dogs such that Ralph believes that they have fleas'. Even supposing (1) and (2) to be in fact de dicto, however, they are quite different in form from 'Ralph believes that some dogs have fleas', as they involve no quantifier phrases. And it has been claimed[13] that sentences such as 'Ralph believes that Cicero is bald' (if not also (2) and more generally all de dicto belief ascriptions whose that-clauses involve only demonstratives, indexicals, and proper names) do imply the corresponding de re ascription. Restricting himself to cases such as (1) Quine, at least in "Quantifiers and Propositional Attitudes,"[14] states that it *is* permissible from it to infer (3), even though his official doctrine in that article makes it hard to see how this is to be accounted for. Clearly, intuitions waver here. David Kaplan's "Quantifying In"[15] not only harbors quite a different intuition, but even offers a general account of why the inference in question must fail. For my part, I approve of his conclusion that (1) *does not* imply (3) (and likewise that (2) does not imply (4)) if—but only if—Frege's assumption is maintained that belief ascriptions such as (1) and (2) can be simultaneously true.

---

[11] See Theodore Sider, "Three Problems for Richard's Theory of Belief Ascription," *Canadian Journal of Philosophy*, xxv, 4 (1995): 487-514.

[12] Two such philosophers are Richard and Graeme Forbes: Richard, "Propositional Attitudes," in Bob Hale and Crispin Wright, eds., *A Companion to the Philosophy of Language* (Cambridge: Blackwell, 1997), pp. 196-226; and Forbes, "Substitutivity and the Coherence of Quantifying In," *The Philosophical Review*, cv, 3, (1996): 337-72. Sider has shown that Richard's theory, as set forth in his book *Propositional Attitudes*, validates the transition from (1) to (3) but not that from (2) to (4)—in "Three Problems for Richard's Theory of Belief Ascription."

[13] Again, by Richard and Forbes, among others.

It is Kripke's result that can help us understand what exactly goes wrong in the argument above. (1) and (2) are assumed to be true. Clearly, in order to make sense of how (3) and (4) (as well as (5) and (6)) could be true, we must evaluate

(7) Ralph believes that $x$ is bald.

with respect to arbitrary assignments, and in particular with respect to some assignment $\alpha$, mapping the free variable $x$ onto Cicero. Now, we have seen above that Kripke's result urges us to distinguish the believer's language from our own (that is, from the language in which belief ascriptions are couched). The occurrences of 'Cicero' and 'Tully', in 'Cicero is bald' and 'Tully is not bald' to which Ralph assents, belong to *his* language. On the other hand, the variable $x$ occurring in (7) is, when taken together with the assignment $\alpha$ that assigns Cicero to $x$, a (temporary) name of Cicero in *our* language. It plays the same role as 'London', as occurring in 'Pierre believes that London is pretty', which definitely belongs to our language. Which names in Ralph's language are we to map (by way of translation) onto the variable $x$? Direct reference theorists are famous for holding, controversially, that proper names have no other semantic role to play except to refer to their bearers. But it was never controversial that this is the only semantic role played by free variables. All that $x$ does, relative to assignment $\alpha$, is to refer to Cicero. How are we then to translate the relevant fragment of Ralph's language into that fragment of our own language which is used in (7)? Clearly, the homophonic translation is the only sensible choice for the predicate 'is bald'. As for 'Cicero' and 'Tully', it seems to me that they are *both* to be mapped onto our $x$, relative to $\alpha$, as there is no reason for invidiously mapping either of them without the other. Now, however, the same argument applies that prompted us to conclude that 'Pierre believes that London is pretty' is neither true nor false. Just as there was no translation from Pierre's language into our own that could do justice to Pierre's rationality, there is no admissible translation from (the relevant fragment of) Ralph's language into (the relevant fragment of) ours. In the absence of any admissible translation, both

---

[14] In *The Ways of Paradox* (New York: Random House, 1966), pp. 183-94, reprinted in Leonard Linsky, ed., *Reference and Modality* (New York: Oxford, 1971), pp. 101-11: "The kind of exportation which leads from (14) to (15) should doubtless be viewed in general as implicative"—(14) and (15) being, respectively, 'Ralph believes that Ortcutt is a spy', and 'Ralph believes $z$($z$ is a spy) of Ortcutt)' (p. 188).

[15] In Donald Davidson and Jaakko Hintikka, eds., *Words and Objections: Essays on the Work of W. V Quine* (Boston: Reidel, 1969), pp. 206-42.

(7) Ralph believes that $x$ is bald.

and

(8) Ralph does not believe that $x$ is bald.

ought to be taken as being neither true nor false with respect to the assignment $\alpha$. This gives us an excellent reason why existential generalization should fail.[16] That is, it fails *if* pairs of sentences such as (1) and (2) above are taken to be simultaneously true. Let us now turn to this assumption.

## III

As I just said, the conclusion we have reached—that existential generalization fails—depends on the assumption that for some proper names $\xi$ and $\zeta$, such that $\xi = \zeta$, both

(a) $S$ believes that...$\xi$...

and

(b) $S$ does not believe that...$\zeta$...

hold. But are we sure that we want to maintain this assumption? Are we really prepared to hold that Ralph, by accepting 'Cicero is bald' while rejecting 'Tully is bald', thereby shows that he believes that Cicero is bald but does not believe that Tully is bald, and must be so described? Let us bear in mind that the statements which Ralph accepts (and those he rejects)—in particular, 'Cicero is bald' and 'Tully is bald', respectively—belong to *his* language. On the other hand, 'Ralph believes that Cicero is bald but does not believe that Tully is' belongs to *our* language. Trivially, in order for the latter to be true, it is necessary that its first conjunct,

---

[16] In my annual tour of Kaplan's "Quantifying In," I was recently surprised to find that in footnote 33 the author says almost as much: "Another way out is to accept the fact that two names may represent the same person to Ralph though Ralph believes the non-identity, but to put an ad hoc restriction on exportation. For example to analyze (33) ['Ralph *Bel* ('$x$ is a spy', Ortcutt)'] as '$\exists\alpha$ [$R(\alpha$, Ortcutt, Ralph) & Ralph $B\ulcorner\alpha$ is a spy$\urcorner$] & $\sim\exists\alpha$ [$R(\alpha$, Ortcutt, Ralph) & $\sim$ Ralph $B\ulcorner\alpha$ is a spy$\urcorner$]'. This prevents exportation where contradiction threatens. But again much that we would like to say is inexpressible in Quine's nomenclature" (p. 242). Unfortunately, Kaplan blames this conclusion for its ad hocness. That it is not at all ad hoc is what "A Puzzle about Belief" shows.

(9) Ralph believes that Cicero is bald.

also be true. Let us restrict our attention to the name 'Cicero', forgetting about 'Tully' for the time being. And let us restrict our language to that fragment of it which only contains 'Cicero' as a name of the Roman orator. Surely, the belief ascription (9), just as any other sentence, is true in the larger language if and only if it is true in any fragment of it in which it is expressible at all. Obvious as it is, this last statement amounts to a general principle which should not go unnoticed. Let us list it along with the others we used in proving Kripke's result:

> *The sublanguage principle* (SP): for any sentence *s* belonging to a language *L*, *s* is true in *L* if and only if it is true in any sublanguage *L'* of *L*, to which it belongs.

It is hard to imagine how this principle could possibly fail to be true: How could either adding to (the *if* direction) or omitting from (the *only if* direction) a language any expression not occurring in statement *s* make any difference as to the truth value of *s* itself? Actually, a much weaker, and correspondingly safer, principle will suffice for our purposes: WSP is just like SP, except that the clause is added "in any sublanguage *L'* of *L* (to which it belongs) which can be obtained from *L* by omitting from it one proper name only."

Now, is 'Ralph believes that Cicero is bald' true in that sublanguage of English that only has the name 'Cicero' to refer to the Roman orator? This question can be answered, as usual, first by noting to which statements Ralph assents, and then by considering how they are to be translated into the (sub)language at hand. We already know that Ralph accepts 'Cicero is bald' and rejects 'Tully is bald'. Now, which names *in Ralph's language* are to be mapped by way of translation onto *our* 'Cicero', belonging to the sublanguage at hand? Clearly, Ralph's 'Cicero' ought to be translated as our 'Cicero'. What about Ralph's 'Tully'? In our larger language, it is translated as 'Tully', of course, which is no longer available in the sublanguage. If we leave it untranslated, then we deprive ourselves of the possibility of taking all the available evidence into account, as we do not know what to make of Ralph's dissent from 'Tully is bald'. Therefore, no other possibility remains, it seems to me, but to translate Ralph's 'Tully' as our 'Cicero'. Now, however, the familiar argument applies once again: the only translation available fails to do justice to Ralph's rationality. A number of belief ascriptions to Ralph—both 'Ralph believes that Cicero is bald', and 'Ralph does not believe that Cicero is bald' and, by parity of reasoning, both 'Ralph believes that Tully is bald' and 'Ralph does not believe that Tully is bald'—all turn out to be neither true nor false. The inescapable conclusion, it seems to me, is that also

(10) Ralph believes that Cicero is bald but does not believe that Tully is bald.

is neither true nor false. That our (larger) language has *two* names, 'Cicero' and 'Tully', homophonic to Ralph's 'Cicero' and 'Tully', makes it appear as though Ralph's case is very different from Pierre's. But, once our attention is drawn to SP, this appearance is shown to be quite misleading: their cases are really completely analogous. They are analogous also in that the position of the believer with respect to the two names involved—'London' and 'Londres', respectively, 'Cicero' and 'Tully'—is entirely symmetric. In a subsequent section, I shall consider other, nonsymmetric, cases in which, for example, Ralph assents to 'Cicero is bald' and has never heard of the name 'Tully'. We shall then see that, in such cases, 'Ralph believes that Cicero is bald' is true and merely adding the name 'Tully' to our language does not make it truth valueless, if Ralph has no disposition as to 'Tully is bald'. (This lends further support to the *if* part of SP above). I shall also consider more difficult cases, such as those in which Ralph does have some negative disposition concerning 'Tully is bald', insufficient though to counterbalance his positive disposition concerning 'Cicero is bald'.

## IV

Frege's assumption, that Ralph's state of mind is accurately represented by (1) and (2) above, has to be rejected. Surely, this is a radical reform of our pretheoretical intuitions. Let us now take stock. Note, first of all, that having accepted SP along with the others, it can no longer happen that the same belief ascription can be true when it is taken as belonging to the believer's language, and neither true nor false when taken as belonging to ours. Consider, for example, 'Ralph believes that Cicero was bald but does not believe that Tully is bald', which was seen to be neither true nor false in our language. Could it be true in Ralph's own language? No, for SP applies to Ralph's language as well. So the first half of that sentence, 'Ralph believes that

Cicero was bald', is neither true nor false in the fragment of his language which contains only 'Cicero'—as the same argument we applied to our own language applies here to Ralph's; similarly for the second half of that sentence.

Second, assume—which we surely must—that Ralph is disposed to assent sincerely to 'Cicero is Cicero'. Of course, the belief ascription (in our language) 'Ralph believes that Cicero is Cicero' is true. Suppose now that Ralph is also disposed—which, given the first part of our story, he surely must be—to assent sincerely to 'Cicero is not Tully'. Given that the homophonic translation into our language is available (unlike in Pierre's case, where only our 'London' was available as a translation of both his 'London' and 'Londres'), the ascription 'Ralph believes that Cicero is not Tully' is true. Can we apply SP here? No, because there is no eligible sublanguage, not containing either 'Cicero' or 'Tully', in which the sentence in question is expressible. Our theory is therefore able to account for the fact that it is perfectly possible for anyone to believe that Cicero is Cicero, without believing that Cicero is Tully. This is an important result, especially in view of the fact that nowhere in our theory were propositions, thoughts, *Sinns*, concepts, and the whole apparatus that goes with them so much as mentioned.

So far, we have seen that the following belief ascriptions:

(a) Ralph believes that Cicero is bald.
(b) Ralph believes that Tully is bald.

together with their negations, are neither true nor false. (Incidentally, even though this goes against the intuitions of almost every philosopher, I am convinced that it complies rather well with the intuitions of the philosophically unsophisticated. If you ask any such person, 'Does Ralph believe that Cicero is bald?', she is likely, I think, to answer, 'Yes and no'—if she can answer at all. Of course, she will not hesitate to admit that Ralph holds opposite views about the two *names* 'Cicero' and 'Tully', namely, that the former refers to a bald person and the latter does not). On the other hand:

(c) Ralph believes that Cicero is Cicero.
(d) Ralph believes that Cicero is not Tully.

are both true. This, too, is as it should be. Can we add any other belief ascription to either list? Having rejected Frege's assumption that Ralph's state of mind was accurately represented by (1) and (2) above (that is, by the truth of (a) and the falsity of (b)), we should reconsider our above rejection of existential generalization, as it rested on the former, now rejected, assumption.

Shall we then say that existential generalization preserves truth in general? No, because there are cases—involving, for example, quantifier phrases and definite descriptions—that are intuitively invalid. Surely, from

(11) Keith suspects the shortest spy is a spy.

one cannot validly infer:

(12) The shortest spy is someone Keith suspects is a spy.

As a matter of fact, our amended theory of belief ascription together with the principles appealed to so far can certify this. In order for us to ascribe a subject $S$ the belief that $p$, it is necessary and sufficient that $S$ assents to some $p^*$ intertranslatable with $p$. Now, translation was assumed to be compositional. In order for it to be true that Keith suspects the shortest spy is a spy—thus making (11) true—it is enough that Keith sincerely assents to something like 'The shortest spy is a spy'. But for it to be true that there is an $x$, who is the shortest spy and is such that Keith suspects $x$ is a spy, Keith must sincerely assent to something like '$N$ is a spy', where '$N$' is translatable into our language as '$x$' (under some assignment mapping it onto the individual referred to by '$N$', if it exists). Now, proper names (and possibly other expressions) can be *compositionally* translated by some free variable, acting as a temporary proper name (relative to some assignment); but definite descriptions cannot. As Keith might be unwilling to assent to any statement of the form '$N$ is a spy'—where '$N$' is a proper name or some other expression that admits of translation as a free variable '$x$'—(12) is seen not to be implied by (11).

But this same argument also shows that existential generalization cannot fail to preserve truth whenever its premise is of the form '$S$ believes that $p$', where the that-clause, 'that $p$', contains no quantifier phrases and no singular terms except proper names.[17] The argument could also show the same for indexicals and demonstratives, provided it could be shown that they, too, are translatable as free variables (relative to some assignment). This would vindicate the intuition of those philosophers, mentioned above, who claim that all de dicto belief ascriptions whose that-clauses involve only demonstratives, indexicals, and proper names do imply the corresponding de re ascriptions. Their intuition cannot be trusted so long as such pairs of belief ascriptions as (1) and (2) are taken to be simultaneously true; but as we have seen, that possibility has to be rejected. My hunch is that that intuition is correct, but to establish it, the point about translation must be thoroughly discussed first—a worthy project, as it is an interesting point in itself, on quite general grounds. (It is fortunate, I think, that more than one set of intuitions is at play here, as those pertaining only to the validity of some inferences and to the characterization of de re belief have conspicuously failed to provide evidence enough to settle the issue conclusively. We can expect more definite conclusions to emerge from the interaction between such intuitions and those pertaining to translation. Of course, we never quite lost sight of the issue of translation, as shown, for example, by the center-stage position held by Rudolf Carnap's theory of belief ascription, where the notion of intensional isomorphism looms large, which is immediately related with translation in general.)

---

[17] The interested reader can verify, as an easy exercise, that from 'Cicero = Tully and Ralph believes that Cicero ≠ Tully' (which turns out to be true, as we have seen above), it is permissible to infer '$\exists x \exists y (x = y \wedge \text{Ralph believes that } x \neq y)$.

I do not intend to explore the issue of translation to any great depth and shall only put forth a few scattered remarks. The main objection to the idea that free variables can translate indexicals and demonstratives goes as follows. Even supposing one grants that proper names can sometimes be translated as free variables (relative to some assignments, assigning the right objects to them), this can only be on the ground that (a) free variables can be taken as temporary names of the objects assigned to them, and (b) proper names have no linguistic meanings that translations are bound to preserve. But precisely for this reason, indexicals and demonstratives cannot be translated as free variables, as they do in fact have linguistic meanings, possibly represented by their characters (in Kaplan's sense), which are entirely missing from free variables. Therefore, just as taking 'Rudolf is clever' to be a translation of 'He [pointing at Rudolf Carnap] is clever' would amount to straining the notion of translation beyond recognition, so, a fortiori, '$x$ is clever' (relative to the assignment of Rudolf Carnap to the variable '$x$') cannot be taken to be a translation of the statement 'He is clever' (accompanied by the pointing)—end of objection. To a certain extent, it might merely be a matter of terminology and of what we take to be the relevant notion of translation. One can, of course, insist that translations must be the kind of word-for-word correlations to be found in dictionaries. Clearly, no dictionary could possibly map the name 'Rudolf' onto the personal pronoun 'he', much less onto any free variable. In real life, however, translations are likely to be a much more unruly business. For one thing, the circumstances in which the translations are taking place must be taken into account. Suppose that, pointing at Rudolf Carnap, someone utters in English, 'He is clever'. A blind person who does not understand English might ask (in Italian, say) for a translation. Were one to offer 'Rudolf è intelligente', would you object that, properly speaking, this was no translation? Or suppose that the President of the United States, visiting Italy, says 'I am grateful for your hospitality': would the translator be wrong in offering 'Il Presidente è grato per l'ospitalità' or 'Egli è grato per l'ospitalità' as a translation? I do not think so. Similarly, the Italian demonstrative 'qui' ('here'), accompanied by a demonstration to a region in the vicinity of the speaker but at some distance from the translator, is best translated by the latter as 'there' ('lì') with the appropriate demonstration (which, incidentally, *must* differ from the original).

It is a much debated issue, though still far from being settled, what it is that proper translation must preserve. Mark Sainsbury,[18] for one, has remarked that room is to be made for "the possibility that tokens of intuitively nonsynonymous types should have the same sense. If on Monday I utter 'It is sunny today', you can on Tuesday report me in the words 'He said that it was sunny yesterday'. In a Fregean spirit, I assume that the relevant part of an accurate report of an utterance has the same sense as the utterance reported. So my token of 'It is sunny today' has the same sense as your token of 'It was sunny yesterday', even though they are of nonsynonymous types" (*ibid.*). Now, is it Fregean sense or is it synonymy (or, possibly, neither) which translation must aim to preserve? I do not intend to probe into this problem any deeper, especially in view of the fact that indexicals and demonstratives are peripheral to our main concern. Let me mention only in passing that, in the end, it might turn out that their references and their syntactic structure are the only things that, not negotiably, the translating words must have in common with the translated ones.

It might be objected that even in some cases where the that-clause involves only proper names, existential generalization is unwarranted. Empty names, in particular, raise a problem. Can one go from

(13) Keith is certain that Sherlock Holmes could solve the case.

to

(14) Sherlock Holmes is someone Keith is certain could solve the case.

and then to

(15) There is someone Keith is certain could solve the case.

I shall be even more brief here. It is not quite obvious to me that the transition is *not* truth preserving in this case. We must bear in mind that the proper name 'Sherlock Holmes', though it occurs within the that-clause of (13), belongs to *our* language; and it is not obvious that we have the right to use a proper name as a proper name, unless we assume—or at least pretend to assume—that there is an individual to which the name refers. A number of authors (among them John McDowell, Gareth Evans, and Kendall Walton[19]) maintain that statements such as (13) are best understood as ones in which the speaker is simply playing along with his subject and pretending that the fictional character exists. Within that pretense, it is surely permissible to conclude that (15) holds, too. This, of course, raises a number of problems, chiefly among them, whether (13), so understood, is true only 'in the pretense' or is just true simpliciter. A related problem is, what truth-value status do such statements as 'Sherlock Holmes = Sherlock Holmes' have? Evans and Walton pointed out that one can make a serious claim about a fiction (and about someone else's beliefs involving fictional characters) by making a pretend assertion within a game of make believe licensed by the fiction. This is a large issue, still wide open and not at all restricted to belief ascription. It is fair to conclude, therefore, that the argument I am considering—to the effect that existential generalization fails to preserve truth in the case of empty names—is less than conclusive.[20] A more in-depth discussion of this issue must be left for another occasion, but it has to be noted that, in any case, it has virtually nothing to do with belief ascription, as it arises in connection with any sentence in which an empty name occurs, whether or not the belief predicate also occurs in it.

---

[18] "Fregean Sense," in Timothy Childers, Petr Kolar, and Vladimir Svoboda, eds., *Logica '96* (Prague: Filosofia, 1997), pp. 261-76.

## V

In Kripke's puzzle, it was crucial that Pierre's attitudes toward the two names, 'London' and 'Londres', were entirely symmetric: both names were just as vivid for him, and he used both as names of London. In addition, his relationship with the city of London was hardly closer after he moved there than it was when he was still living in France (or so can the example be modified). Similarly for Ralph, with respect to 'Cicero' and 'Tully' and the famous Roman orator. Let us now consider situations that are not as symmetric. Let Ralph assent to the statement that Mount Everest is snowy, and suppose he never comes across the name 'Gaurisanker', so that, were the question 'Is Gaurisanker snowy?' put to him, he would surely answer 'I don't know'. What do our principles tell us here? Shall we say that the ascription 'Ralph believes that Mount Everest is snowy' is true or that it is neither true nor false? And what about 'Ralph believes that Mount Gaurisanker is snowy'?

---

[19] Evans, *The Varieties of Reference*, McDowell, ed. (New York: Oxford, 1982); McDowell, "On the Sense and Reference of a Proper Name," reprinted in A. W. Moore, ed., *Meaning and Reference* (New York: Oxford, 1993), pp. 111-36; Walton, *Mimesis as Make-Believe: On the Foundation of the Representational Arts* (Cambridge: Harvard, 1990).

[20] A different strategy for refusing to take the transition from (13) to (14) as a counterexample to existential generalization is to point out that the name 'Sherlock Holmes' changes its reference in going from (13) to (14). Forbes takes this view: "My view, then, is that those who hold that [{(13), (14)}] to constitute a counterexample to exportation are interpreting 'Sherlock Holmes' as denoting an actual fictional character in [(13)] and as an empty name, purportedly of a real individual, in [(14)]. This switch deprives the alleged counterexample of any force" (*op. cit.*, p. 360).

It is not quite clear to me whether a name a speaker has never heard of ought to be taken as belonging to his language at all. If 'Gaurisanker' does not belong to Ralph's language, then it is not clear that his answer to our question, 'Is Gaurisanker snowy?', is even remotely relevant to the issue of ascribing him beliefs. (Carl Gustav Hempel was known, within a restricted circle of friends, by the name 'Peter'. Surely, it does not count against a philosopher who knows that Hempel was a leader of the logical empiricist movement that she cannot answer the question 'Was Peter a leader of the logical empiricist movement?') But let us change our example a little. Suppose that Ralph *is* to some extent familiar with that name—suppose, for example, that he has heard it being used by others, or that he has used it himself, possibly in asking questions like 'Where is Gaurisanker?', or even in sincerely asserting 'Gaurisanker is a mountain, somewhere'. What shall we say? In general, the statements assented to (or dissented from) by a speaker ought to be considered as evidence for ascribing beliefs to him. This is not the only evidence relevant to the issue of belief ascription, as other behavioral evidence, not amounting to either assent or dissent, should, in general, also be considered. But linguistic evidence has a very special role to play whenever belief is involved. Its privileged status is due to the fact that assertion is intrinsically, if indirectly, related with belief—as shown by Moore's paradox[21]—in a way that nonlinguistic behavior is not. It then follows that what a person says, which statements she assents to and which she dissents from, are all important in establishing what her beliefs are. But is everything she says to be taken at face value? Quite generally, not all the available evidence is to be given the same weight in assessing a given hypothesis. Belief ascription is no exception in this respect. Just as, in ascribing some degree of, say, temperature to a given object we are likely simply to ignore the results of some measurements that cannot be squared with the results given by other, more reliable, ones, so we are entitled to ignore some of Ralph's statements, either on the ground of his poor command of the name 'Gaurisanker' or—more crucially—on the grounds that, in some circumstances, even Ralph himself is not a reliable source of information concerning his own beliefs. We are all familiar with some instruments of measurement not being trustworthy in some circumstances—for example, when the items to be measured are not presented in the proper way. Think of how any optical device or, for that matter, our own visual apparatus, is subject to malfunction in poor light conditions. More generally, any instrument is likely to give incorrect results both in cases where the incoming data are insufficiently detailed and when excess noise cannot be screened off. Similarly, we must take people as not equally reliable sources of information concerning their own beliefs in all circumstances. This is to say that their utterances, playing a role comparable to that of the readings of measuring devices, are not to be taken at face value in every case.

---

[21] I intend to explore this issue in another article.

The analogy between people and measuring devices is, I think, quite close. Incidentally, it shows what is wrong with the idea, cherished by some philosophers, of accounting for behavior like Ralph's in terms of modes of presentation and that belief is really a three-place relation, relating persons, propositions, and modes of presentation. In regarding belief as such, these philosophers represent a person's utterances as infallibly revealing what that person believes. Whenever Ralph, for example, both assents to 'Mount Everest is snowy' and puzzles over 'Mount Gaurisanker is snowy', they account for his behavior by describing Ralph as believing Mount Everest to be snowy under some mode of presentation (somehow related with the name 'Everest') and at the same time not believing that Mount Everest, that is, Gaurisanker, is snowy under some other mode of presentation (related with 'Gaurisanker'). Everything that Ralph says and does is thus taken into account and is given as much weight as possible. The price they must pay, however, is jettisoning the ordinary notion of belief and inadvertently substituting it with the quite different notion of belief-under-a-mode-of-presentation. The analogue of this move for measuring devices would be regarding a measuring device as being invariably reliable by tinkering with the magnitude it is supposed to be measuring. Consider a police speedometer. Instead of admitting that its reliability is subject to severe constraints, one could take it to be measuring the speed of a car under particular modes of presentation—at dusk, at night, in heavy traffic, and so on.

Surely, this would turn the speedometer into an infallible device, but one whose use was no longer clear. I must leave this large topic for some other article to consider.

Let us go back to Ralph, Mount Everest, and Gaurisanker. Suppose, having actually climbed Mount Everest, Ralph confidently asserts 'Mount Everest is snowy'. Assuming that it makes good sense to talk of degrees of belief, surely we want to say that Ralph believes at the maximum degree (1) that mount Everest is snowy. Later on, however, when the question is put to him: 'Is Mount Gaurisanker snowy?', he either answers 'I don't know' or even, possibly, 'I don't think so'. Either way, and for whatever reason, his position is definitely asymmetric with respect to the two names. Now, the upshot of the discussion so far is that we are entitled to regard his answers involving the name 'Gaurisanker' as unreliable, that is, not revelatory of his true beliefs. We are therefore entitled not only to take him as believing that Everest is snowy, but even as believing that Gaurisanker is snowy. What is more, we can even take him as believing it to degree 1.

So far, in this section, we have just been clarifying our intuitions. Let us now see if the AT of belief ascription can account for them. Ignoring all of Ralph's utterances involving the name 'Gaurisanker' is tantamount to taking his language to be restricted to that fragment of it with only the name 'Everest' to refer to the highest mountain on Earth, as we do not have to translate the rest into our language. Now, the name 'Everest' can be equally translated into our language either as 'Everest' or as 'Gaurisanker'. Choosing the latter on AT makes 'Ralph believes that Gaurisanker is snowy' true. And choosing the former makes 'Ralph believes that Everest is snowy' true. This is to say that—restricted to this sort of case—the principle of substitutivity, in the form:

> *Substitutivity*: every conditional of the form $a=b \rightarrow (S \rightarrow S')$—where $a$ and $b$ are proper names, and $S'$ is obtained from $S$ by substituting one or more occurrences of $a$ with occurrences of $b$—is true.

holds according to our theory. That principle also holds with respect to ascriptions such as 'Ralph believes that Cicero is bald'—the substitution of 'Cicero' with 'Tully' being permissible here, as it takes us from an ascription that is neither true nor false to another that is likewise neither true nor false. We have seen above, however, that 'Ralph believes that Cicero $\neq$ Tully' is true whereas 'Ralph believes that Cicero $\neq$ Cicero' is not, as this would amount to representing Ralph as irrational—which, presumably, he is not. Substitutivity therefore does not hold in general. What does generally hold is, rather, a different principle—let us call it *Leibniz's law*—licensing the *permutation* of co-referring proper names.

> *Leibniz's law*: every conditional of the form $a = b \rightarrow (S \rightarrow S')$—where $a$ and $b$ are proper names, and $S'$ is obtained from $S$ by simultaneously substituting every occurrence of $a$ with an occurrence of $b$, and every occurrence of $b$ with an occurrences of $a$—is true.

It is easy to see why Leibniz's law holds. Let Ralph assent to the statement 'Cicero ≠ Tully'. Surely, the translation is available that maps 'Cicero' in Ralph's language onto 'Cicero' in our language, and similarly his 'Tully' onto our 'Tully', thus making 'Ralph believes that Cicero ≠ Tully' true (SP does not apply here). But another translation is available (satisfying the usual principles) that maps his 'Cicero' onto our 'Tully' and his 'Tully' onto our 'Cicero', thus making 'Ralph believes that Tully ≠ Cicero' true too. I omit a more formal proof of this result, but the general idea should be clear enough.[22] Once the basic intuitions are straightened out, all the jigsaw pieces seem to fit together quite nicely and belief ascriptions turn out not to require any substantial amendment of the basic laws of logic. Senses, modes of presentation, propositions, and related entities were left entirely out of the picture.

MARCO SANTAMBROGIO

Università di Parma

---

[22] It would be interesting to compare the form of Leibniz's law given in the text with that given by Richard (*op. cit.*, p 199). It seems to me that the former avoids some unwelcome consequence of the latter, but a full discussion of this point must be left to another article.

## BOOK REVIEWS

*Descartes Embodied: Reading Cartesian Philosophy through Cartesian Science.* DANIEL GARBER. New York: Cambridge University Press, 2001. xii + 337 p. Cloth $64.95, paper $22.95.

This book is a collection of fourteen previously published articles, half of which appeared in the 1980s, and half in the 1990s. During this time, Daniel Garber has been in the vanguard of those transforming the quality of the history of philosophy in the Anglophone world, and these essays exemplify the meticulous scholarship and historical sensitivity that we have now come to expect from those working in the early modern period. To these virtues, Garber adds some special ones of his own: an unfailing clarity of style, subtlety, and ingenuity in argument, and a refusal to get bogged down in detail for its own sake. Garber understands what his successors beavering away in the field now often lose sight of, that minute textual analysis is only philosophically interesting when we keep in view the (philosophical, scientific, cultural) backdrop that brings the detailed debates into relief.

The subtitle of the book makes clear Garber's guiding idea in the collection as a whole. He implicitly follows the interpretive rule made famous by the great Cartesian scholar Charles Adam[1]—that Descartes's metaphysics should be seen as primarily motivated by his scientific aspirations. The book as a whole provides a persuasive alternative to the epistemological lens through which the Cartesian system tended to be viewed through much of the twentieth century; thus, in a well-known paper entitled "Semel in Vita," Garber aims to "set the epistemological project of the *Meditations* into the broader context of the Cartesian program for science, and show why Descartes thought that such an epistemological project was a necessary preliminary to scientific investigation" (223). It is not that Descartes has no interest in the nature of knowledge, the refutation of skepticism, and so on. But his main aim is to establish the "groundwork for his revolution in physics, and for the arguments that establish the world of mechanism and allow us to set aside the commonsense and sense-bound world that the Aristotelians have mistaken for the world we live in" (255-56).

---

[1] Cf. Adam and Paul Tannery, eds., *Oeuvres de Descartes* (Paris: Cerf, 1897-1913), Volume XII, p. 143.

0022-362X/02/9912/648–50

In the Introduction, Garber sets out his own credo as a historian of philosophy by contrasting his stance with that of Bernard Williams.[2] For Williams, although the history of philosophy must "constitute its object, the work, in genuinely historical terms," there is nevertheless a "cut-off point, where authenticity is replaced as the objective by the aim of articulating philosophical ideas" (4, quoting from Williams's celebrated study of Descartes). To which Garber replies that, if we are to learn philosophy from Descartes, as opposed to using him as a mere foil for our contemporary views, then we must "try to reach genuine understanding of what he thinks." And this is an exercise that will often require us to "go beyond the texts themselves and into the contemporary culture to understand their presuppositions" (5).

The main bulwark of seventeenth-century intellectual culture that Garber highlights again and again in these studies is the still dominant scholasticism against which Descartes fought. Many of the essays are devoted to expounding and illuminating the Cartesian approach to explaining the physical universe—an approach which resolutely bypasses the "substantial forms" and "real qualities" so beloved of Descartes's scholastic predecessors. Those interested in the concept of causation will find much food for thought here, partly because Cartesian dualism (positing the existence of incorporeal souls apparently possessed of causal powers to affect matter) seems to introduce a joker into the pack as far as the laws of Cartesian physics are concerned, and partly because, if Garber's general line is right, it implies that Descartes's view of how the physical universe works turns out to be rather less "modern" and "scientific" than we are often inclined to suppose.

In this connection, Garber offers a very thorough discussion of G. W. Leibniz's criticism that in Descartes's system, mind-body interaction would violate the laws of conservation: no use to say that the soul merely alters the *direction* of motion of neural impulses, not the overall quantity of motion, since (Leibniz argues) a properly formulated conservation law will require any change (whether of speed or direction) to be compensated for by appropriate changes in other bodies. Garber argues that Descartes himself did not, in fact, propose the "change of direction" maneuver (though his "fountain keeper" analogy in the *Traité de l'homme* seems to me to come pretty close to it); but that, in any case, Descartes was quite happy with the idea that minds are able to cause bodily changes that violate conservation: "even though finite, imperfect minds may act in some law like way

---

[2] *Descartes: The Project of Pure Enquiry* (Atlantic Highlands, NJ: Humanities, 1978).

...the motion *they* cause need not satisfy the conservation principle.... *They* may add or subtract motion from the world, even if *God* cannot.... The conservation principle governs only *purely material* systems in nature, systems in which God is the *only* cause of motion" (165). There is no space here to discuss whether this interpretation of Descartes is defensible, but even if it is, this is one place where one may perhaps hanker for Williams's "cut-off point." For the position imputed to Descartes seems so bizarre (implying, so far as I can see, that humans could produce indefinite amounts of free energy) that it is hard to see how it could be more than a historical curiosity, unable to mesh with any of our contemporary philosophical concerns about the mind and its place in nature.

One of the later papers, "Descartes and Occasionalism," goes some way to dispelling this worry, by showing how Descartes's views on the three possible kinds of causal interaction (body-body, mind-body, and body-mind) can be located right in the middle of an intriguing developmental story in the history of ideas—one which begins with the schoolmen's attribution of causal powers to natural objects (in virtue of their substantial forms), and runs through to the later occasionalist view (found in Malebranche) that only an infinitely perfect being can ground the necessary connection between cause and effect (a view that, paradoxically, paves the way for David Hume's banishing of the concept of connection from the domain of knowable science). Reflecting on the way in which this story has continued up to our own day (something which Garber, perfectly reasonably, forbears to do) may lead the reader to wonder whether the standard description of Descartes as the father of modernity is really very apt. Notwithstanding the apparent circularity of many of the explanations they proffered, is not the schoolmen's idea that things have "basic, inborn tendencies to behavior" (206) rather closer to capturing our modern conception of causal powers (the powers found certainly in the organic world, and arguably even at the molecular and atomic level) than is the notion of ghostly thrusting, which the Cartesian worldview seems to attribute to divine and even human agents? However that may be, this is a splendid and thought-provoking collection of papers which all philosophers interested in the early modern period will be glad to have on their shelves.

JOHN COTTINGHAM

University of Reading

# NOTES AND NEWS

The Editors regret the death of David Lewis, Professor of Philosophy at Princeton University. He died on October 14, 2001 at the age of 60. Born in Oberlin, Ohio, Lewis received a degree in chemistry from Swarthmore College. While spending a term at Oxford University, he became enthused with philosophy, which he continued to study on returning to the United States. After graduating in 1964, Lewis took a Ph.D. at Harvard University under the supervision of W.V. Quine. In 1966, he became Assistant Professor at the University of California, Los Angeles, Associate Professor at Princeton in 1970, and full Professor there three years later. Beginning in 1971, Lewis spent the northern summer in Australia no fewer than twenty-seven times. As a result, his influence on Australian philosophy was immense. His works include: *Convention* (1969), *Counterfactuals* (1973), *On The Plurality Of Worlds* (1986), *Parts of Classes* (1991), and five volumes of philosophical papers (1983, 1986, 1998, 1999, 2000). His work spanned all the major areas of philosophy, setting the agenda for contemporary debate in metaphysics and epistemology, in philosophical logic, the philosophy of language, the philosophy of mind, in value theory and ethics, and in the philosophy of mathematics.

Among these many contributions, Lewis is perhaps most celebrated for his "modal realism," according to which non-actual possible worlds are not just an heuristic device for the organization of intuitions about possibility and necessity, but entities as real and concrete as the actual universe. Lewis's vision was broad, but his primary focus was metaphysical. The guiding idea, which Lewis termed "Humean supervenience," was the thesis that there are no necessary connections, that "all there is to the world is a vast mosaic of local matters of particular fact." The program was to account for everything "else" that there manifestly is in the world, within the bounds of this austere constraint.

The brilliance is in the details: a formal semantics for the counterfactual conditional from comparative similarity of worlds; a subtle treatment of the interplay of time and chance; the rehabilitated counterfactual put to work in an analysis of causation; mental states accounted for as the occupants of causal roles, and so on. Taken separately, each of these little parts of the mosaic is a gem in its own right; together they constitute a great philosophical achievement.

# INDEX TO VOLUME XCIX

## INDEX OF SUBJECTS DISCUSSED IN THE JOURNAL

**Action, Philosophy of**

McCarty, Richard: The Maxims Problem. 99 (1): 29-44 '02

**Akrasia**

Arpaly, Nomy: Moral Worth. 99 (5): 223-45 '02

**Antirealism**

Achinstein, Peter: Is There a Valid Experimental Argument for Scientific Realism? 99 (9): 470-95 '02

**Aristotle**

Keeley, Brian L.: Making Sense of the Senses: Individuating Modalities in Humans and Other Animals. 99 (1): 5-28 '02

**Azzouni, Jody**

Katz, Jerrold J.: Mathematics and Metaphilosophy. 99 (7): 362-90 '02

**Belief**

Santambrogio, Marco: Belief and Translation. 99 (12): 624-47 '02

**Benacerraf, Paul**

Katz, Jerrold J.: Mathematics and Metaphilosophy. 99 (7): 362-90 '02

**Big Bang**

Bostrom, Nick: Self-locating Belief in Big Worlds: Cosmology's Missing Link to Observation. 99 (12): 607-23 '02

**Biology, Philosophy of**

Ariew, André: Two Ways of Thinking about Fitness and Natural Selection. By Mohan Matthen and André Ariew 99 (2): 55-83 '02

Keeley, Brian L.: Making Sense of the Senses: Individuating Modalities in Humans and Other Animals. 99 (1): 5-28 '02

Matthen, Mohan: Two Ways of Thinking about Fitness and Natural Selection. By Mohan Matthen and André Ariew. 99 (2): 55-83 '02

**Bonjour, Lawrence**

Olsson, Erik J.: What Is the Problem of Coherence and Truth? 99 (5): 246-72 '02

**Carnap, Rudolf**

Murdoch, Dugald: Induction, Hume, and Probability. 99 (4): 185-99 '02

Restall, Greg: Carnap's Tolerance, Meaning, and Logical Pluralism. 99 (8): 426-43 '02

**Causation**

McDermott, Michael: Causation: Influence versus Sufficiency. 99 (2): 84-101 '02

Pereboom, Derk: Robust Nonreductive Materialism. 99 (10): 499-531 '02

Yablo, Stephen: De Facto Dependence. 99 (3): 130-48 '02

**Cohen, G. A.**

Carter, Alan: Value-pluralist Egalitarianism. 99 (11): 577-99 '02

**Coherence**

Olsson, Erik J.: What Is the Problem of Coherence and Truth? 99 (5): 246-72 '02

**Compatibilism**

Widerker, David: Farewell to the Direct Argument. 99 (6) 316-24 '02

**Consciousness**

Block, Ned: The Harder Problem of Consciousness. 99 (8): 391-425 '02

**Cosmology**

Bostrom, Nick: Self-locating Belief in Big Worlds: Cosmology's Missing Link to Observation. 99 (12): 607-23 '02

**Counterfactuals**

McDermott, Michael: Causation: Influence versus Sufficiency. **99** (2): 84-101 '02

Yablo, Stephen: De Facto Dependence. **99** (3): 130-48 '02

**De Facto Dependence**

Yablo, Stephen: De Facto Dependence. **99** (3): 130-48 '02

**Democracy**

McMahon, Christopher: Why There Is No Issue between Habermas and Rawls. **99** (3): 111-29 '02

**Determinism**

Widerker, David: Farewell to the Direct Argument. **99** (6) 316-24 '02

**Dilemma, Prisoner's**

Ullmann-Margalit, Edna: Trust out of Distrust. **99** (10) 532-48 '02

**Dworkin, Ronald**

Carter, Alan: Value-pluralist Egalitarianism. **99** (11): 577-99 '02

Roemer, John E.: Egalitarianism against the Veil of Ignorance. **99** (4): 167-84 '02

**Egalitarianism**

Carter, Alan: Value-pluralist Egalitarianism. **99** (11): 577-99 '02

Roemer, John E.: Egalitarianism against the Veil of Ignorance. **99** (4): 167-84 '02

**Epistemology:** *See* Knowledge, Theory of

**Equality**

Carter, Alan: Value-Pluralist Egalitarianism. **99** (11): 577-99 '02

**Ethics**

Arpaly, Nomy: Moral Worth. **99** (5): 223-45 '02

McCarty, Richard: The Maxims Problem. **99** (1): 29-44 '02

McMahon, Christopher: Why There Is No Issue between Habermas and Rawls. **99** (3): 111-29 '02

Widerker, David: Farewell to the Direct Argument. **99** (6) 316-24 '02

Zimmerman, Michael J.: Taking Luck Seriously. **99** (11): 553-76 '02

**Evolution**

Keeley, Brian L.: Making Sense of the Senses: Individuating Modalities in Humans and Other Animals. **99** (1): 5-28 '02

**Evolutionary Theory**

Ariew, André: Two Ways of Thinking about Fitness and Natural Selection. By Mohan Matthen and André Ariew. **99** (2): 55-83 '02

Matthen, Mohan: Two Ways of Thinking about Fitness and Natural Selection. By Mohan Matthen and André Ariew. **99** (2): 55-83 '02

**Fitness**

Ariew, André: Two Ways of Thinking about Fitness and Natural Selection. By Mohan Matthen and André Ariew. **99** (2): 55-83 '02

Matthen, Mohan: Two Ways of Thinking about Fitness and Natural Selection. By Mohan Matthen and André Ariew. **99** (2): 55-83 '02

**Free Will**

McCarty, Richard: The Maxims Problem. **99** (1): 29-44 '02

Zimmerman, Michael J.: Taking Luck Seriously. **99** (11): 553-76 '02

**Frege, Gottlob**

Santambrogio, Marco: Belief and Translation. **99** (12): 624-47 '02

**Functionalism**

Pereboom, Derk: Robust Nonreductive Materialism. **99** (10): 499-531 '02

**Game Theory**

Ullmann-Margalit, Edna: Trust out of Distrust. **99** (10) 532-48 '02

**Gödel, Kurt**

Katz, Jerrold J.: Mathematics and Metaphilosophy. 99 (7): 362-90 '02

**Habermas, Jürgen**

McMahon, Christopher: Why There Is No Issue between Habermas and Rawls. 99 (3): 111-29 '02

**Harsanyi, John**

Roemer, John E: Egalitarianism against the Veil of Ignorance. 99 (4). 167-84 '02

**Hobbes, Thomas**

Ullmann-Margalit, Edna: Trust out of Distrust 99 (10) 532-48 '02

**Hume, David**

Murdoch, Dugald: Induction, Hume, and Probability. 99 (4): 185-99 '02

**Impartiality**

Roemer, John E.: Egalitarianism against the Veil of Ignorance. 99 (4): 167-84 '02

Sen, Amartya: Open and Closed Impartiality. 99 (9): 445-69 '02

**Induction**

Murdoch, Dugald: Induction, Hume, and Probability. 99 (4): 185-99 '02

**Justice**

McMahon, Christopher: Why There Is No Issue between Habermas and Rawls. 99 (3): 111-29 '02

Sen, Amartya: Open and Closed Impartiality. 99 (9): 445-69 '02

**Kant, Immanuel**

Arpaly, Nomy: Moral Worth. 99 (5)· 223-45 '02

**Kim, Jaegwon**

Pereboom, Derk: Robust Nonreductive Materialism 99 (10): 499-531 '02

**Kitcher, Philip**

Ariew, André: Two Ways of Thinking about Fitness and Natural Selection By Mohan Matthen and André Ariew. 99 (2): 55-83 '02

Matthen, Mohan: Two Ways of Thinking about Fitness and Natural Selection. By Mohan Matthen and André Ariew. 99 (2). 55-83 '02

**Knowledge, Theory of**

Block, Ned: The Harder Problem of Consciousness. 99 (8): 391-425 '02

Koethe, John: Stanley and Williamson on Knowing How. 99 (6): 325-28 '02

Schiffer, Stephen: Amazing Knowledge. 99 (4): 200-02 '02

**Kripke, Saul**

Santambrogio, Marco: Belief and Translation. 99 (12)· 624-47 '02

**Language, Philosophy of**

Restall, Greg: Carnap's Tolerance, Meaning, and Logical Pluralism. 99 (8): 426-43 '02

Santambrogio, Marco: Belief and Translation. 99 (12): 624-47 '02

Sider, Theodore: The Ersatz Pluriverse. 99 (6): 279-315 '02

**Lewis, C. I.**

Olsson, Erik J.. What Is the Problem of Coherence and Truth? 99 (5)· 246-72 '02

**Lewis, David**

McDermott, Michael: Causation: Influence versus Sufficiency. 99 (2): 84-101 '02

Sider, Theodore: The Ersatz Pluriverse. 99 (6): 279-315 '02

Yablo, Stephen: De Facto Dependence. 99 (3): 130-48 '02

**Logic, Philosophy of**

Restall, Greg: Carnap's Tolerance, Meaning, and Logical Pluralism. 99 (8): 426-43 '02

**Luck**

Zimmerman, Michael J.: Taking Luck Seriously. 99 (11): 553-76 '02

**Materialism**

Pereboom, Derk: Robust Nonreductive Materialism. 99 (10): 499-531 '02

**Mathematics, Philosophy of**

Katz, Jerrold J.: Mathematics and Metaphilosophy. 99 (7): 362-90 '02

**Maxims Problem**

McCarty, Richard: The Maxims Problem. 99 (1): 29-44 '02

**Meaning**

Restall, Greg: Carnap's Tolerance, Meaning, and Logical Pluralism. 99 (8): 426-43 '02

**Memo to Authors**

99 (1): 54 '02

**Metaphilosophy**

Katz, Jerrold J.: Mathematics and Metaphilosophy. 99 (7): 362-90 '02

**Metaphysics**

Morreau, Michael: What Vague Objects Are Like. 99 (7): 333-61 '02

**Mind, Philosophy of**

Block, Ned: The Harder Problem of Consciousness. 99 (8): 391-425 '02

Pereboom, Derk: Robust Nonreductive Materialism. 99 (10): 499-531 '02

**Moral Worth**

Arpaly, Nomy: Moral Worth. 99 (5): 223-45 '02

**Multiple Realizability Thesis**

Pereboom, Derk: Robust Nonreductive Materialism. 99 (10): 499-531 '02

**Nagel, Thomas**

Zimmerman, Michael J.: Taking Luck Seriously. 99 (11): 553-76 '02

**Naturalism**

Block, Ned: The Harder Problem of Consciousness. 99 (8): 391-425 '02

**Natural Selection**

Ariew, André: Two Ways of Thinking about Fitness and Natural Selection. By Mohan Matthen and André Ariew. 99 (2): 55-83 '02

Matthen, Mohan: Two Ways of Thinking about Fitness and Natural Selection. By Mohan Matthen and André Ariew. 99 (2): 55-83 '02

**New Books**

99 (4) 210-22 '02

99 (10) 549-54 '02

**New Books: Anthologies**

99 (5) 273-78 '02

**New Books: Translations**

99 (2): 109-10 '02

**Newton, Isaac**

Ariew, André: Two Ways of Thinking about Fitness and Natural Selection. By Mohan Matthen and André Ariew. 99 (2): 55-83 '02

Matthen, Mohan: Two Ways of Thinking about Fitness and Natural Selection. By Mohan Matthen and André Ariew. 99 (2): 55-83 '02

**Notes and News**

99 (7) 390 '02

99 (8) 444 '02

99 (9) 498 '02

99 (10) 552 '02

99 (12) 651 '02

**Observation**

Bostrom, Nick: Self-locating Belief in Big Worlds: Cosmology's Missing Link to Observation. 99 (12): 607-23 '02

**Ontology**

Morreau, Michael: What Vague Objects Are Like. 99 (7): 333-61 '02

**Perception**

Keeley, Brian L.: Making Sense of the Senses: Individuating Modalities in Humans and Other Animals. 99 (1)· 5-28 '02

**Perrin, Jean**

Achinstein, Peter: Is There a Valid Experimental Argument for Scientific Realism? 99 (9)· 470-95 '02

**Philosophy, Moral:** *See* Ethics

**Philosophy, Political:** *See* Political Theory

**Physics**

Bostrom, Nick: Self-locating Belief in Big Worlds: Cosmology's Missing Link to Observation 99 (12): 607-23 '02

**Physics, Philosophy of**

Bostrom, Nick: Self-locating Belief in Big Worlds: Cosmology's Missing Link to Observation. 99 (12): 607-23 '02

**Political Theory**

Carter, Alan: Value-pluralist Egalitarianism. 99 (11): 577-99 '02

McMahon, Christopher: Why There Is No Issue between Habermas and Rawls. 99 (3): 111-29 '02

Roemer, John E.: Egalitarianism against the Veil of Ignorance. 99 (4): 167-84 '02

Sen, Amartya: Open and Closed Impartiality. 99 (9): 445-69 '02

Ullmann-Margalit, Edna: Trust out of Distrust. 99 (10) 532-48 '02

**Possible Worlds**

Sider, Theodore: The Ersatz Pluriverse. 99 (6): 279-315 '02

**Prisoner's Dilemma:** *See* Dilemma, Prisoner's

**Probability**

Murdoch, Dugald: Induction, Hume, and Probability. 99 (4): 185-99 '02

**Qualia**

Keeley, Brian L: Making Sense of the Senses: Individuating Modalities in Humans and Other Animals. 99 (1): 5-28 '02

**Quine, W. V.**

Katz, Jerrold J.: Mathematics and Metaphilosophy. 99 (7): 362-90 '02

Restall, Greg: Carnap's Tolerance, Meaning, and Logical Pluralism. 99 (8): 426-43 '02

Santambrogio, Marco: Belief and Translation. 99 (12): 624-47 '02

**Rationalism**

Katz, Jerrold J.: Mathematics and Metaphilosophy. 99 (7): 362-90 '02

**Rawls, John**

Carter, Alan: Value-pluralist Egalitarianism. 99 (11). 577-99 '02

McMahon, Christopher: Why There Is No Issue between Habermas and Rawls. 99 (3): 111-29 '02

Roemer, John E.: Egalitarianism against the Veil of Ignorance. 99 (4): 167-84 '02

Sen, Amartya: Open and Closed Impartiality. 99 (9): 445-69 '02

**Realism**

Achinstein, Peter: Is There a Valid Experimental Argument for Scientific Realism? 99 (9): 470-95 '02

Block, Ned: The Harder Problem of Consciousness. 99 (8): 391-425 '02

Katz, Jerrold J.: Mathematics and Metaphilosophy. 99 (7): 362-90 '02

**Responsibility**

Widerker, David: Farewell to the Direct Argument. 99 (6) 316-24 '02

Zimmerman, Michael J.: Taking Luck Seriously. 99 (11): 553-76 '02

**Rousseau, Jean-Jacques**

Ullmann-Margalit, Edna: Trust out of Distrust. 99 (10) 532-48 '02

**Ryle, Gilbert**

Koethe, John: Stanley and Williamson on Knowing How. 99 (6): 325-28 '02

Schiffer, Stephen: Amazing Knowledge. 99 (4): 200-02 '02

**Salmon, Wesley C.**

Achinstein, Peter: Is There a Valid Experimental Argument for Scientific Realism? 99 (9): 470-95 '02

**Scanlon, Thomas**

Sen, Amartya: Open and Closed Impartiality. 99 (9): 445-69 '02

**Science, Philosophy of**

Achinstein, Peter: Is There a Valid Experimental Argument for Scientific Realism? 99 (9): 470-95 '02

Ariew, André, Two Ways of Thinking about Fitness and Natural Selection. By Mohan Matthen and André Ariew. 99 (2): 55-83 '02

Bostrom, Nick: Self-locating Belief in Big Worlds: Cosmology's Missing Link to Observation. 99 (12): 607-23 '02

Keeley, Brian L.: Making Sense of the Senses: Individuating Modalities in Humans and Other Animals. 99 (1): 5-28 '02

Matthen, Mohan: Two Ways of Thinking about Fitness and Natural Selection. By Mohan Matthen and André Ariew. 99 (2): 55-83 '02

**Sen, Amartya**

Carter, Alan: Value-pluralist Egalitarianism. 99 (11): 577-99 '02

**Senses**

Keeley, Brian L.: Making Sense of the Senses: Individuating Modalities in Humans and Other Animals. 99 (1): 5-28 '02

**Smith, Adam**

Sen, Amartya: Open and Closed Impartiality. 99 (9): 445-69 '02

**Sober, Elliott**

Ariew, André: Two Ways of Thinking about Fitness and Natural Selection. By Mohan Matthen and André Ariew. 99 (2): 55-83 '02

Matthen, Mohan· Two Ways of Thinking about Fitness and Natural Selection. By Mohan Matthen and André Ariew. 99 (2): 55-83 '02

**Translation**

Santambrogio, Marco: Belief and Translation. 99 (12): 624-47 '02

**Trust**

Ullmann-Margalit, Edna· Trust out of Distrust. 99 (10) 532-48 '02

**Truth**

Olsson, Erik J.: What Is the Problem of Coherence and Truth? 99 (5): 246-72 '02

Sider, Theodore: The Ersatz Pluriverse. 99 (6): 279-315 '02

**Vagueness**

Morreau, Michael: What Vague Objects Are Like. 99 (7): 333-61 '02

**Van Fraassen, Bas**

Achinstein, Peter: Is There a Valid Experimental Argument for Scientific Realism? 99 (9): 470-95 '02

**Williams, Bernard**

Zimmerman, Michael J.: Taking Luck Seriously. 99 (11): 553-76 '02

**Wittgenstein, Ludwig**

Katz, Jerrold J.: Mathematics and Metaphilosophy. 99 (7): 362-90 '02

# INDEX TO VOLUME XCIX

## INDEX OF AUTHORS

Including Contributing Authors and Reviewers

**Achinstein, Peter,** Is There a Valid Experimental Argument for Scientific Realism? 99 (9): 470-95 '02

**Allen, James,** Review of Wolfgang-Rainer Mann: *The Discovery of Things: Aristotle's* Categories *and Their Context*. 99 (6): 329-32 '02

**Ariew, André,** Two Ways of Thinking about Fitness and Natural Selection. By Mohan Matthen and André Ariew. 99 (2): 55-83 '02

**Arpaly, Nomy,** Moral Worth. 99 (5): 223-45 '02

**Block, Ned,** The Harder Problem of Consciousness. 99 (8): 391-425 '02

**Bostrom, Nick,** Self-locating Belief in Big Worlds: Cosmology's Missing Link to Observation. 99 (12): 607-23 '02

**Byrd, David,** Review of Ruth Garrett Millikan: *On Clear and Confused Ideas: An Essay about Substance Concepts*. By Robert Cummins, Alexa Lee, Martin Roth, David Byrd, and Pierre Poirier. 99 (2). 102-08 '02

**Carter, Alan,** Value-pluralist Egalitarianism. 99 (11): 577-99 '02

**Cottingham, John,** Review of Daniel Garber: *Descartes Embodied: Reading Cartesian Philosophy through Cartesian Science*. 99 (12): 648-50 '02

**Cummins, Robert,** Review of Ruth Garrett Millikan: *On Clear and Confused Ideas: An Essay about Substance Concepts*. By Robert Cummins, Alexa Lee, Martin Roth, David Byrd, and Pierre Poirier. 99 (2): 102-08 '02

**Darwall, Stephen,** Review of John Rawls: *Lectures on the History of Moral Philosophy*. 99 (1). 49-53 '02

**Engstrom, Stephen,** Review of Allen Wood: *Kant's Ethical Thought*. 99 (3): 149-52 '02

**Freeman, Samuel,** Review of Brian Barry: *Culture and Equality*. 99 (11): 600-06 '02

**Haldane, John,** Review of Peter Geach: *Truth and Hope*. 99 (3): 157-62 '02

**Katz, Jerrold J.,** Mathematics and Metaphilosophy. 99 (7): 362-90 '02

**Keeley, Brian L.,** Making Sense of the Senses: Individuating Modalities in Humans and Other Animals. 99 (1): 5-28 '02

**Koethe, John,** Stanley and Williamson on Knowing How. 99 (6): 325-28 '02

**Lee, Alexa,** Review of Ruth Garrett Millikan: *On Clear and Confused Ideas: An Essay about Substance Concepts*. By Robert Cummins, Alexa Lee, Martin Roth, David Byrd, and Pierre Poirier. 99 (2): 102-08 '02

**Lovibond, Sabina,** Review of Richard Wollheim: *On the Emotions*. 99 (3): 153-56 '02

**Matthen, Mohan,** Two Ways of Thinking about Fitness and Natural Selection. By Mohan Matthen and André Ariew 99 (2): 55-83 '02

**McCarty, Richard,** The Maxims Problem. 99 (1): 29-44 '02

**McDermott, Michael,** Causation: Influence versus Sufficiency. 99 (2): 84-101 '02

**McMahon, Christopher,** Why There Is No Issue between Habermas and Rawls. 99 (3): 111-29 '02

**Meinwald, Constance C.,** Review of Richard Sorabji: *Emotion and Peace of Mind: From Stoic Agitation to Christian Temptation*. 99 (3): 163-66 '02

**Morreau, Michael,** What Vague Objects Are Like. 99 (7): 333-61 '02

**Murdoch, Dugald,** Induction, Hume, and Probability. 99 (4): 185-99 '02

**Olsson, Erik J.,** What Is the Problem of Coherence and Truth? 99 (5): 246-72 '02

**Pereboom, Derk,** Robust Nonreductive Materialism. 99 (10): 499-531 '02
**Poirier, Pierre,** Review of Ruth Garrett Millikan: *On Clear and Confused Ideas: An Essay about Substance Concepts* By Robert Cummins, Alexa Lee, Martin Roth, David Byrd, and Pierre Poirier. 99 (2): 102-08 '02

**Restall, Greg,** Carnap's Tolerance, Meaning, and Logical Pluralism. 99 (8): 426-43 '02
**Roemer, John E.,** Egalitarianism against the Veil of Ignorance. 99 (4): 167-84 '02
**Roth, Martin,** Review of Ruth Garrett Millikan: *On Clear and Confused Ideas: An Essay about Substance Concepts.* By Robert Cummins, Alexa Lee, Martin Roth, David Byrd, and Pierre Poirier. 99 (2): 102-08 '02

**Santambrogio, Marco,** Belief and Translation. 99 (12): 624-47 '02
**Schiffer, Stephen,** Amazing Knowledge. 99 (4): 200-02 '02
**Sen, Amartya,** Open and Closed Impartiality. 99 (9): 445-69 '02
**Sider, Theodore,** Review of Lynne Rudder Baker: *Persons and Bodies. A Constitution View.* 99 (1): 45-48 '02
The Ersatz Pluriverse. 99 (6): 279-315 '02

**Tierney, Richard,** Review of Deborah K. W. Modrak: *Aristotle's Theory of Language and Meaning.* 99 (4) 203-09 '02

**Ullmann-Margalit, Edna,** Trust out of Distrust. 99 (10) 532-48 '02

**Widerker, David,** Farewell to the Direct Argument. 99 (6) 316-24 '02

**Yablo, Stephen,** De Facto Dependence. 99 (3): 130-48 '02

**Zimmerman, Michael J.,** Taking Luck Seriously. 99 (11): 553-76 '02

## INDEX OF BOOKS REVIEWED

Listed by Author of Item Reviewed

**Baker, Lynne Rudder:** *Persons and Bodies: A Constitution View*—Theodore Sider. 99 (1): 45-48 '02
**Barry, Brian:** *Culture and Equality*—Samuel Freeman. 99 (11). 600-06 '02
**Garber, Daniel:** *Descartes Embodied: Reading Cartesian Philosophy through Cartesian Science*—John Cottingham. 99 (12): 648-50 '02
**Geach, Peter:** *Truth and Hope*—John Haldane. 99 (3): 157-62 '02
**Mann, Wolfgang-Rainer:** *The Discovery of Things: Aristotle's* Categories *and Their Context*—James Allen. 99 (6): 329-32 '02
**Millikan, Ruth Garrett:** *On Clear and Confused Ideas: An Essay about Substance Concepts*—Robert Cummins, Alexa Lee, Martin Roth, David Byrd, and Pierre Poirier. 99 (2): 102-08 '02
**Modrak, Deborah K. W.:** *Aristotle's Theory of Language and Meaning*—Richard Tierney. 99 (4) 203-09 '02
**Rawls, John:** *Lectures on the History of Moral Philosophy*—Stephen Darwall. 99 (1): 49-53 '02
**Sorabji, Richard:** *Emotion and Peace of Mind. From Stoic Agitation to Christian Temptation*—Constance C. Meinwald. 99 (3): 163-66 '02
**Wollheim, Richard:** *On the Emotions*—Sabina Lovibond. 99 (3): 153-56 '02
**Wood, Allen:** *Kant's Ethical Thought*—Stephen Engstrom. 99 (3) 149-52 '02

United States Postal Service

## Statement of Ownership, Management, and Circulation

1. Publication Title: THE JOURNAL OF PHILOSOPHY

2. Publication Number: 0 0 2 2 - 3 6 2 X

3. Filing Date: October 16, 2002

4. Issue Frequency: MONTHLY

5. Number of Issues Published Annually: 12

6. Annual Subscription Price: Individual $35.00 / Institution $75.00

7. Complete Mailing Address of Known Office of Publication (Not printer) (Street, city, county, state, and ZIP+4)

The Journal of Philosophy
1150 Amsterdam Avenue, Mail Code 4972 New York NY 10027

Contact Person: Pam Ward

Telephone: 212-666-4419

8. Complete Mailing Address of Headquarters or General Business Office of Publisher (Not printer)

SAME

9. Full Names and Complete Mailing Addresses of Publisher, Editor, and Managing Editor (Do not leave blank)

Publisher (Name and complete mailing address)

The Journal of Philosophy
1150 Amsterdam Avenue, Mail Code 4972
New York NY 10027

Editor (Name and complete mailing address)

Arthur Danto, President
The Journal of Philosophy, Inc.
1150 Amsterdam Avenue, Mail Code 4972, New York NY 10027

Managing Editor (Name and complete mailing address)

John Smylie, Managing Editor
The Journal of Philosophy
1150 Amsterdam Avenue, Mail Code 4972 New York NY 10027

10. Owner (Do not leave blank. If the publication is owned by a corporation, give the name and address of the corporation immediately followed by the names and addresses of all stockholders owning or holding 1 percent or more of the total amount of stock. If not owned by a corporation, give the names and addresses of the individual owners. If owned by a partnership or other unincorporated firm, give its name and address as well as those of each individual owner. If the publication is published by a nonprofit organization, give its name and address.)

| Full Name | Complete Mailing Address |
|---|---|
| THE JOURNAL OF PHILOSOPHY, INC. | 1150 Amsterdam Avenue |
| | Mail Code 4972 |
| | New York New York 10027 |
| | |
| | |

11. Known Bondholders, Mortgagees, and Other Security Holders Owning or Holding 1 Percent or More of Total Amount of Bonds, Mortgages, or Other Securities. If none, check box → ☒ None

| Full Name | Complete Mailing Address |
|---|---|
| | |
| | |
| | |
| | |
| | |

12. Tax Status (For completion by nonprofit organizations authorized to mail at nonprofit rates) (Check one)
The purpose, function, and nonprofit status of this organization and the exempt status for federal income tax purposes:
☒ Has Not Changed During Preceding 12 Months
☐ Has Changed During Preceding 12 Months (Publisher must submit explanation of change with this statement)

13. Publication Title: THE JOURNAL OF PHILOSOPHY

14. Issue Date for Circulation Data Below: DECEMBER 2002

| 15. Extent and Nature of Circulation | | Average No. Copies Each Issue During Preceding 12 Months | No. Copies of Single Issue Published Nearest to Filing Date |
|---|---|---|---|
| a. Total Number of Copies (Net press run) | | 3500 | 3500 |
| b. Paid and/or Requested Circulation | (1) Paid/Requested Outside-County Mail Subscriptions Stated on Form 3541 (Include advertiser's proof and exchange copies) | 700 | 666 |
| | (2) Paid In-County Subscriptions Stated on Form 3541 (Include advertiser's proof and exchange copies) | 2400 | 2220 |
| | (3) Sales Through Dealers and Carriers, Street Vendors, Counter Sales, and Other Non-USPS Paid Distribution | | |
| | (4) Other Classes Mailed Through the USPS | | |
| c. Total Paid and/or Requested Circulation (Sum of 15b. (1), (2), (3) and (4)) ▶ | | 3100 | 2886 |
| d. Free Distribution by Mail (Samples, complimentary, and other free) | (1) Outside-County as Stated on Form 3541 | | |
| | (2) In-County as Stated on Form 3541 | | |
| | (3) Other Classes Mailed Through the USPS | | |
| e. Free Distribution Outside the Mail (Carriers or other means) | | 150 | 150 |
| f. Total Free Distribution (Sum of 15d. and 15e.) ▶ | | 150 | 150 |
| g. Total Distribution (Sum of 15c. and 15f) ▶ | | 3250 | 3036 |
| h. Copies not Distributed | | 250 | 464 |
| i. Total (Sum of 15g. and h.) ▶ | | 3500 | 3500 |
| j. Percent Paid and/or Requested Circulation (15c. divided by 15g. times 100) | | 95.38 | 95.06 |

16. Publication of Statement of Ownership
☑ Publication required. Will be printed in the December 2002 issue of this publication. ☐ Publication not required.

17. Signature and Title of Editor, Publisher, Business Manager, or Owner

John Smylie, Managing Editor

Date: 11/19/02

I certify that all information furnished on this form is true and complete. I understand that anyone who furnishes false or misleading information on this form or who omits material or information requested on the form may be subject to criminal sanctions (including fines and imprisonment) and/or civil sanctions (including civil penalties).

The Center for Philosophy of Religion at the University of Notre Dame announces five fellowships for the 2003-2004 academic year two Postdoctoral Fellowships (stipend of $35,000 teaching one course each semester, otherwise $30,000), a Distinguished Scholar Fellowship (stipend of $50,000 teaching one course each semester; otherwise $45,000), an Extraordinary Fellowship ($40,000) to a scholar who might not qualify for the fellowships above, but who would benefit from a year at the Center—for example, foreign scholars or those outside the field of philosophy (e.g., theologians who want to enrich their understanding of philosophy of religion and theistic philosophy, humanists from other fields who could benefit by a deeper grasp of the sort of philosophy practiced at the Center, and scholars interested in a theistic approach to psychology, sociology, economics and other subjects), and a Visiting Graduate Fellowship ($14,000) to a philosophy graduate student at another institution who is working on a dissertation in philosophy of religion or Christian philosophy and who would profit from spending a year at the Center.

Applicants should see that the Center receives each of the following:

(1) A complete curriculum vitae, which should include a list of publications and names of three people you will ask to send letters of recommendation on your behalf.
(2) Three letters of recommendation.
(3) A statement of no more than three pages (double-spaced) describing the Fellowship for which you are applying and the project on which you would like to work while at the Center.
(4) One published or unpublished paper.

All material should be sent to: Thomas P. Flint, Director, Center for Philosophy of Religion, 418 Malloy Hall, Notre Dame, IN 46556.

The deadline for Extraordinary Fellowship and Distinguished Scholar Fellowship applications for 2003-2004 is January 15, 2003; for all other Fellowships, April 1, 2003.

E-mail: cprelig1@nd.edu    Web site: http://www.nd.edu/~cprelig/

THE JOURNAL OF
PHILOSOPHY

The Journal is pleased to announce to our readers and subscribers the launch of its new website, which will have the pages listed below. The address is: *www.journalofphilosophy.org* or simply *www.jphil.org*.

~ • ~

PURPOSE & HISTORY
CONTACTS AND ADDRESSES
BOARD OF TRUSTEES
EDITORIAL BOARD
MEMOS TO AUTHORS
SUBSCRIPTIONS & ADVERTISING
RECENT & FORTHCOMING ISSUES
ELECTRONIC ISSUES
INDEX 1904–PRESENT
RIGHTS AND PERMISSIONS
LECTURES & CONFERENCES
PHILOSOPHY LINKS

~ • ~

## Dilemmas of Reconciliation

### Cases and Concepts

*Carol A.L. Prager and Trudy Govier, editors*

Paper $34.95 • ISBN 0-88920-416-2

Can enemies who have inflicted unspeakable acts of cruelty on each other live together in peace? This book explores the possibility of forgiveness.

"A brilliant mix of theoretical analysis and empirical case studies that brings out clearly and concisely the political, legal, economic, and psycho-social dimensions of post-conflict reconciliation."
– Howard Adelman, York University

## Structural Idealism

### A Theory of Social and Historical Explanation

*Douglas Mann*

Cloth $42.95 • ISBN 0-88920-391-1

Mann challenges our perception of how cultures and ideals are formed, and shows that these seemingly unalterable structures of our society can be changed, and are changed, by individuals.

## Persons—What Philosophers Say About You

### 2d edition

*Warren Bourgeois*

Paper $29.95 • ISBN 0-88920-379-2

This second edition is redesigned as an introduction to the philosophy of mind and its history, concentrating on current issues such as abortion, euthanasia, genetic engineering, and human rights.

Visit our website at www.wlupress.wlu.ca

WLU Wilfrid Laurier University Press

# Philosophy from Yale

## One World

*The Ethics of Globalization*

Peter Singer

"Many people have written about the economic meaning of globalization; in *One World* Peter Singer explains its moral meaning. His position is carefully developed, his tone is moderate, but his conclusions are radical and profound. **No political theorist or moral philosopher, no public official or political activist, can afford to ignore his arguments**"
—Michael Walzer $21.95

## Renaissance and Reformation

*The Intellectual Genesis*

Anthony Levi

**This wide-ranging book** examines the development of intellectual culture across the breadth of Europe between about 1350 and 1550 $39.95

## Key Concepts in Chinese Philosophy

Zhang Dainian

Translated and edited by Edmund Ryden

"**A valuable contribution** to our understanding of Chinese philosophy by an important scholar of philosophy in China."—Carine Defoort

Culture and Civilization of China Series $45.00

## Balanced Discourses

*A Bilingual Edition*

Xu Gan

English translation by John Makeham
Introductions by Dang Shengyuan and John Makeham

Offers a unique contemporary account of the demise of the Han order.

"**A judicious and eminently readable translation of an important work.**"
—Wai-yee Li

Culture and Civilization of China Series $45.00

## The Elusiveness of the Ordinary

*Studies in the Possibility of Philosophy*

Stanley Rosen

"If adjectives like 'original' and 'unique' had comparative and superlative forms, it would be possible to say that this book is **very original and most unique, as well as insightful and stimulating.** It is also accessible, free of jargon, and rhetorically full-blooded without ever being strident."—Robert B. Pippin $37.50

## Absorbing Perfections

*Kabbalah and Interpretation*

Moshe Idel

Foreword by Harold Bloom

"[This book] is certainly [Idel's] most important volume so far. . . . His work has become **essential for anyone who desires** to confront **Kabbalah.**"
—Harold Bloom, from the Foreword

$45.00

## The Crisis of Reason

*European Thought, 1848-1914*

J. W. Burrow

"So **sweeping** in conception, so persuasive in execution, and, simply, so well written. . . . **An excellent book.**"
—John Banville, *New York Review of Books*

Yale Intellectual History of the West Series

24 illus. $16.95 paperback

## Hegel's Philosophy of Freedom

Paul Franco

"**This is the best book on Hegel to appear in recent years.**"
—Steven B. Smith $22.50 paperback

Yale University Press
yalebooks.com

# new books from the mit press

## Religion and Rationality

**Essays on Reason, God and Modernity**
*Jürgen Habermas*

This important new volume brings together Jürgen Habermas's key writings on religion and religious belief.

Studies in Contemporary German Social Thought series
192 pp. $19.95 paper

## Brain-Wise

**Studies in Neurophilosophy**
*Patricia Smith Churchland*

"In an age of increasing specialization, the gap between philosophy and neuroscience can be enormous. Somehow Churchland manages to connect the dots, and a deeper, richer and more complete view of how the brain enables the mind emerges from her powerful new book. It is a great contribution."
— Michael S. Gazzaniga,
Dartmouth College

A Bradford Book
438 pp., 111 illus. $25 paper

## Imagine There's No Woman

**Ethics and Sublimation**
*Joan Copjec*

A psychoanalytic and philosophical exploration of sublimation as a key term in Jacques Lacan's theories of ethics and feminine sexuality.

264 pp., 14 illus. $29.95

*now in paperback*

## Architectures of Time

**Toward a Theory of the Event in Modernist Culture**
*Sanford Kwinter*

Offers a critical guide to the modern history of time and to the interplay between the physical sciences and the arts.

251 pp., 41 illus. $17.95 paper

## Balkan as Metaphor

**Between Globalization and Fragmentation**
*edited by Dušan I. Bjelić and Obrad Savić*

"This volume's greatest attraction is the way it so wonderfully illustrates the Balkans' baroque complexity, to some the region's insurmountable handicap and to others, including myself, its inimitable charm." — Maria Todorova, University of Illinois at Urbana-Champaign, author of *Imagining the Balkans*

360 pp. $29.95

## The Big Book of Concepts

*Gregory L. Murphy*

"Should be read by anyone who wants to understand the theories and empirical studies that will provide the foundation for new research for decades to come."
— Arthur B. Markman,
University of Texas at Austin

A Bradford Book • 560 pp., 16 illus. $45

## The Letters of George Santayana

**Book Four, 1928–1932**
**The Works of George Santayana**
**Volume V, Book Four**
*George Santayana*
*edited and with an introduction by William G. Holzberger*

The fourth of eight books of the correspondence of George Santayana.

600 pp. $60

*now in paperback*

## Infinity and Perspective

*Karsten Harries*

"Harries's well-positioned and clearly presented history of ideas will generate a great deal of controversy among a select audience of scholars."
— *Publishers Weekly*

392 pp., 18 illus., 6 in color $22.95 paper

To order call 800-405-1619
Prices subject to change without notice

## Essays on Nonconceptual Content

*edited by York H. Gunther*

Offers some of the most important work on nonconceptual content in the philosophy of mind and psychology.

A Bradford Book
384 pp., 3 illus. $30 paper

## Being No One

**The Self-Model Theory of Subjectivity**

*Thomas Metzinger*

A representationalist, empirically based philosophical exploration of consciousness, the subjective experience of selfhood, and the first-person perspective.

A Bradford Book • 584 pp., 14 illus. $55

## Words without Meaning

*Christopher Gauker*

"Essential reading for anyone interested in the nature of linguistic communication." — Marga Reimer, University of Arizona

Contemporary Philosophical Monographs series
A Bradford Book • 312 pp. $25 paper

## Foundations of Cognitive Psychology

**Core Readings**

*edited by Daniel J. Levitin*

An anthology of core readings on cognitive psychology

A Bradford Book
832 pp., 189 illus. $50 paper

*back in print*
*with a new preface and a new chapter*

## Biblical Games

**Game Theory and the Hebrew Bible**

*Steven J. Brams*

"Fascinating. . . logical and deeply rooted in its sources." — *New York Times*

224 pp. $17.95 paper

*now in paperback*

## Justice and Equity

*Serge-Christophe Kolm*
*translated by Harold F. See*
*with the assistance of Denise Killebrew, Chantal Philippon-Daniel, and Myron Rigsby*

"A rich and very important contribution to welfare economics and political philosophy." — Ashley Piggins, *The Economic Journal* (Great Britain)

275 pp., 40 illus. $21.95 paper

*now in paperback*

## Modern Theories of Justice

*Serge-Christophe Kolm*

"A stimulating contribution to the literature." — K. Binmore, *Journal of Economics*

535 pp. $39.95 paper

*now in paperback*

## The Linguistic Turn in Hermeneutic Philosophy

*Cristina Lafont*
*translated by José Medina*

"Lafont's scholarship is consistently first-rate. She provides a superb reconstruction of the development of Habermas's views of language, perhaps the most difficult part of his philosophical theory." — Pablo De Greiff, SUNY Buffalo

Studies in Contemporary German Social Thought
399 pp. $29.95 paper

**ZONE BOOKS**

*now in paperback*

## Remnants of Auschwitz

**The Witness and the Archive**

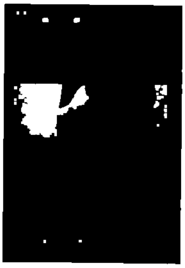
*Giorgio Agamben*
*translated by Daniel Heller-Roazen*

"Sophisticated and insightful." — *Association of Jewish Libraries Newsletter*

Distributed for Zone Books
176 pp. $16 paper

**http://mitpress.mit.edu**

New from
Oxford



**Philosophy and Democracy**
An Anthology
Edited by THOMAS CHRISTIANO
This volume collects some of the leading essays in contemporary democratic theory published in the past thirty years. The chapters address many issues that are central to philosophical reflections on democracy, such as questions pertaining to deliberative and economic approaches
paper $29 95 cloth $65 00

*New in paperback!*
**Morality and Cultural Differences**
JOHN W COOK
Carefully examining the arguments for and against moral relativism, Cook exposes not only that anthropologists have failed in their attempt to support relativism with evidence of cultural differences, but that moral absolutists have been equally unsuccessful in their attempts to refute it.
paper $19 95 cloth $35 00

**Conceivability and Possibility**
Edited by TAMAR SZABO GENDLER and JOHN HAWTHORNE
The capacity to represent things to ourselves as possible plays a crucial role both in everyday thinking and in philosophical reasoning. This volume offers much-needed philosophical illumination of conceivability, possibility, and the relations between them.
paper $29 95 cloth $72 00

*New in paperback!*
**Radical Enlightenment**
Philosophy and the Making of Modernity 1650–1750
JONATHAN I. ISRAEL
*"[A] magnificent ... study of the impact of Spinoza and his philosophy on European cultural history at the hinge of the 17th and 18th centuries .. Sumptuous in the energy, clarity and breadth of its scholarship."*—Los Angeles Times Book Review.
paper $24 95 cloth $45 00

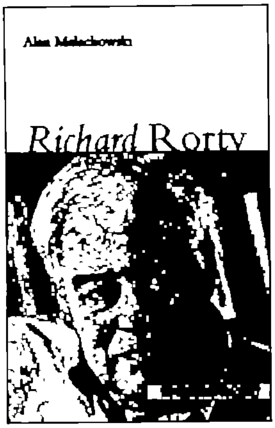
**How Free Are You?**
The Determinism Problem
Second Edition
TED HONDERICH
*"The arguments for free will and determinism are lucidly laid out . A primer that is serviceable, enjoyable and rather mischievous."*—The Observer
*"Refreshing, provocative and original work."*—Times Literary Supplement
paper $19 95

**Thomas Aquinas**
Contemporary Philosophical Perspectives
Edited by BRIAN DAVIES
Brian Davies has collected in one volume the best recent essays on Aquinas by some of the world's foremost scholars of medieval philosophy. Taken together, they illuminate the entire spectrum of Aquinas's thought.
paper $29 95 cloth $65 00

Prices are subject to change and apply only in the US To order, or for more information, please call 1-800-451-7556
Visit our website at www.oup-usa.org

OXFORD

New from Princeton



**Richard Rorty**
**Alan Malachowski**

Richard Rorty is one of the world's most influential living thinkers. Here Alan Malachowski provides an authoritative overview of Rorty's considerable body of work and a general assessment of his impact both within philosophy and in the humanities more broadly.

"This is an engaging, timely, and useful book. It offers the student and general reader something that is badly needed: a wide-ranging, accessible and sympathetic introduction to Rorty's thought."—Michael Williams, Johns Hopkins University

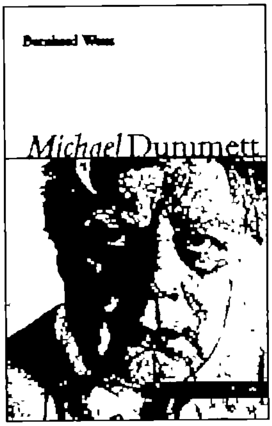
*Philosophy Now Series*
Paper $17.95 ISBN 0-691-05708-7 Cloth $55.00 ISBN 0-691-05707-9
*Available from Princeton only in the Americas*

**Michael Dummett**
**Bernhard Weiss**

No one seriously interested in the philosophy of language can afford to ignore the work of Michael Dummett. This book offers a systematic presentation of the main elements of Dummett's pivotal contribution to contemporary philosophy. Its overarching theme is his discussion of realism: Bernhard Weiss explores the philosopher's characterization of realism, his attack on realism, and his invention and exploration of the anti-realist position.

This book is a sympathetic and accessible study, aiming not only to expose but to engage both with Dummett's philosophical thought and with his philosophical character.

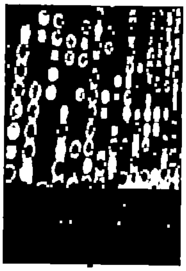
*Philosophy Now Series*
Paper $18.95 ISBN 0-691-11330-0 Cloth $55.00 ISBN 0-691-11329-7
*Available from Princeton only in the Americas*

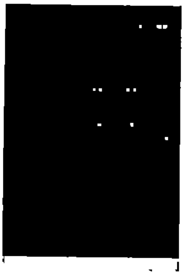


**Princeton University Press**

New from

**Oxford**





Prices are subject to change and apply only in the US To order, or for more information, please call 1-800-451-7556. Visit our website at www oup-usa org

OXFORD

*New in paperback!*

**Nietzsche's Ethics and his War on "Morality"**

SIMON MAY

Nietzsche famously attacked traditional morality, and propounded a controversial ethics of "life-enhancement." Simon May presents a radically new view of Nietzsche's thought, which is shown to be both revolutionary and conservative, and to have much to offer us today after the demise of old values and the "death of God"

paper $19 95 cloth $55 00

**The Oxford Handbook of Epistemology**

Edited by PAUL K. MOSER

The *Oxford Handbook of Epistemology* contains 19 previously unpublished chapters by today's leading figures in the field. These chapters function not only as a survey of key areas, but as original scholarship on a range of vital topics

$74 00

**World Without Design**

The Ontological Consequences of Naturalism

MICHAEL C REA

Philosophical naturalism has dominated the Western academy for well over a century. Here, Michael Rea argues compellingly to the surprising conclusion that naturalists are committed to rejecting realism about material objects, materialism, and perhaps realism about other minds.

$35 00

**Moral Writings**

H. A. PRICHARD

Edited by JIM MACADAM

This is the definitive collection of the ethical work of the great Oxford moral philosopher H A. Prichard (1871–1947). The essays previously included in the posthumous collection *Moral Obligation* are now augmented by a selection of previously unpublished writings from Prichard's manuscripts.

paper $21 95 cloth $65 00

**The Right and the Good**

**Second Edition**

DAVID ROSS

Edited by PHILIP STRATTON-LAKE

Ross's *The Right and the Good* is the pinnacle of ethical intuitionism, which was the dominant moral theory in British philosophy for much of the nineteenth and early twentieth century Stratton-Lake provides the context for a proper understanding of Ross's great work today.

paper $18 95 cloth $39 95

**Consciousness**

New Philosophical Essays

Edited by QUENTIN SMITH and ALEKSANDAR JOKIC

Consciousness is perhaps the most puzzling problem we humans face in trying to understand ourselves. Here, eighteen essays offer new angles on the subject. The contributors discuss such central topics as intentionality, phenomenal content, and the relevance of quantum mechanics to the study of consciousness.

paper $24.95 cloth $74 00